Hewlett Packard
Enterprise

HPE Performance Cluster Manager Installation Guide

Abstract

This publication describes how to install and configure the HPE Performance Cluster Manager 1.3.1 software on an HPE cluster system.

Part Number: 007-6501-008
Published: May 2020
Edition: 2

© Copyright 2018, 2020 Hewlett Packard Enterprise Development LP

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

Arm® is a registered trademark of Arm Limited (or its subsidiaries) in the US and/or elsewhere.

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

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

Oracle® is a registered trademark of Oracle and/or its affiliates.

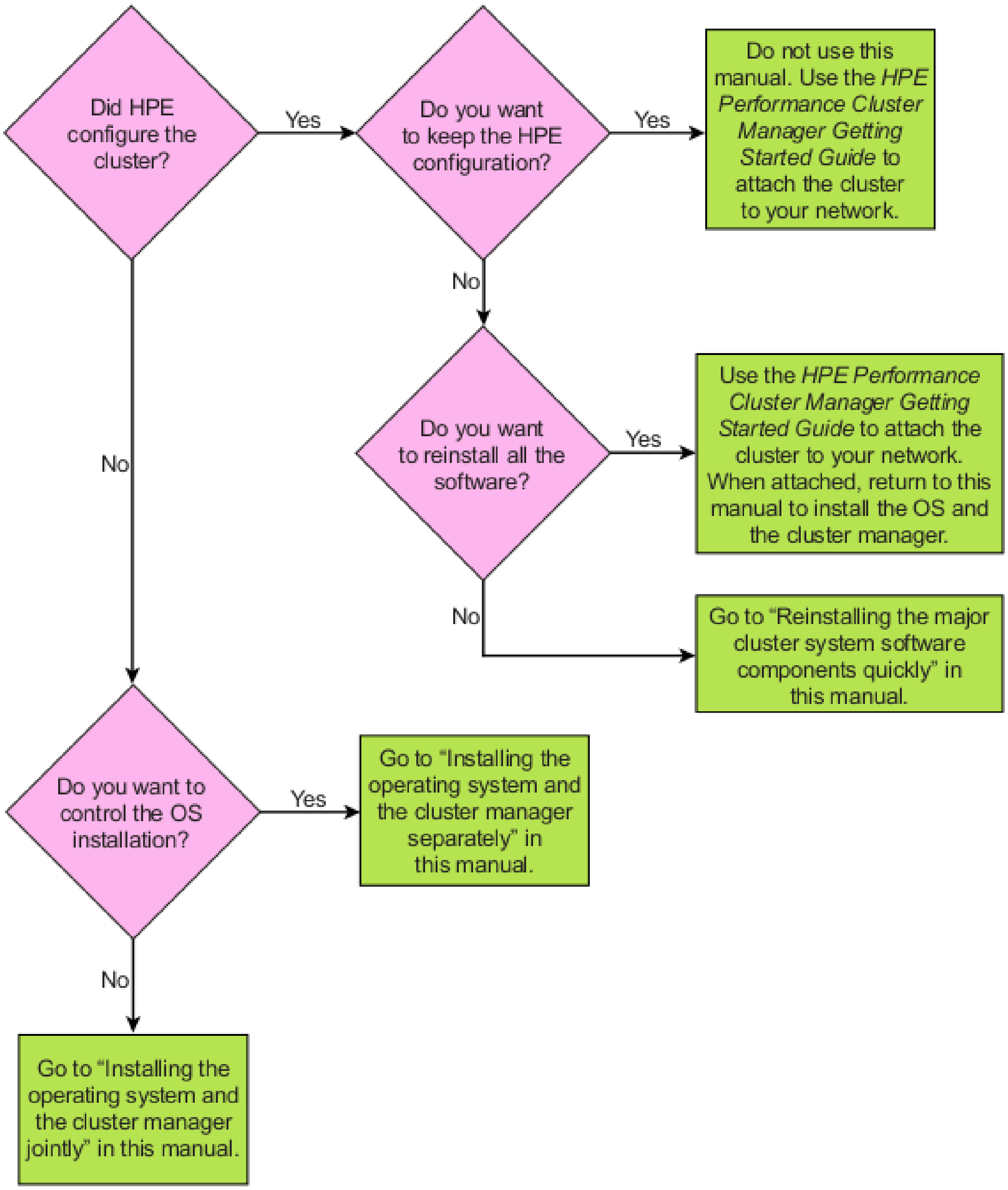
Red Hat® and Red Hat Enterprise Linux® are registered trademarks of Red Hat, Inc., in the United States and other countries.

Contents

Installing HPE Performance Cluster Manager

This manual is written for system administrators, data center administrators, and software developers. The procedures assume that you are familiar with Linux, clusters, and system administration.

The following figure provides an overview of how to proceed with an installation:



Did HPE configure the cluster?
Yes
Do you want to keep the HPE configuration?
Yes
Do not use this manual. Use the *HPE Performance Cluster Manager Getting Started Guide* to attach the cluster to your network.
No
Do you want to reinstall all the software?
Yes
Use the *HPE Performance Cluster Manager Getting Started Guide* to attach the cluster to your network. When attached, return to this manual to install the OS and the cluster manager.
No
No
Go to "Reinstalling the major cluster system software components quickly" in this manual.
Do you want to control the OS installation?
Yes
Go to "Installing the operating system and the cluster manager separately" in this manual.
No
Go to "Installing the operating system and the cluster manager jointly" in this manual.

Figure 1: Installing the cluster manager

The following links direct you to the installation starting points in the preceding figure:

- **Reinstalling the major cluster system software components quickly**
- **Installing the operating system and the cluster manager separately**

- **Installing the operating system and the cluster manager jointly**
- HPE installs the operating system software and the cluster manager software on some cluster systems. If HPE installed and configured the operating system and the cluster software, and you want to keep the configuration, use the procedure in the following manual to attach the cluster to your network:

  HPE Performance Cluster Manager Getting Started Guide

  After you attach the cluster to your site network, you can return to this manual to reconfigure the cluster or add optional features.

The following additional documentation might be useful to you:

- The release notes. The cluster manager release notes include information about features, software packages, and supported platforms. Follow the links on the following website to access the release notes:

  https://www.hpe.com/software/hpcm

  On the product media, the release notes appear in a text file in the following directory:

  `/docs`

  After installation, the release notes and other product documentation reside on the system in the following directory:

  `/usr/share/doc/packages/cm`
- **HPE Performance Cluster Manager Getting Started Guide**
- **HPE Performance Cluster Manager Administration Guide**
- **HPE Performance Cluster Manager Power Management Guide**

NOTE: Before you install RHEL 8.X, check the following website, and make sure that the cluster includes only supported hardware:

https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/8/html/considerations_in_adopting_rhel_8/hardware-enablement_considerations-in-adopting-rhel-8

If your cluster includes a high availability admin node, also note the supported SAS cards at the following website:

https://access.redhat.com/solutions/4444321

Operating system releases supported in the HPE Performance Cluster Manager 1.3.1 release

Obtain operating system software for the cluster directly from your operating system vendor. After you obtain the operating system software, write the `.iso` file to a DVD or USB device. The cluster manager installation instructions assume that the operating system software is written to physical media. If you install the cluster manager from a network location, use your site practices to access the operating system software installation files.

The following table shows the releases that the HPE Performance Cluster Manager 1.3.1 release supports:

Operating system	**Architecture**	**Releases**
SUSE Linux Enterprise Server (SLES)	x86_64 and Arm (AArch64)	SLES 15 SP1, SLES 12 SP5, SLES 12 SP4
Red Hat Enterprise Linux (RHEL)	x86_64	RHEL 8.1, RHEL 8.0, RHEL 7.8, RHEL 7.7
	Arm (AArch64)	RHEL 8.1, RHEL 8.0
CentOS	x86_64	CentOS 8.1, CentOS 8.0, CentOS 7.7

NOTE: The following information pertains to operating system support in this cluster manager release:

- The cluster manager supports RHEL 7.8 as an upgrade from RHEL 7.7 only, not as an initial installation.
- On Arm (AArch64) architecture clusters, cluster manager support for RHEL 8.1 and SLES 12 SP5 is deferred.

  For operating system availability information, see the HPE Support Center page for the HPE Performance Cluster Manager. Click **https://support.hpe.com** and search for **HPCM**.

Within one cluster, nodes can be of a single architecture type or can be a mix of x86_64 and Arm (AArch64) architectures. In mixed-architecture clusters, the admin node and leader nodes (if any) must be x86_64 servers. For HPE Apollo 9000 and HPE SGI 8600 clusters, the admin node must be an HPE ProLiant DL360 server.

You can configure admin nodes for high availability (HA). If you want to configure an HA admin node, purchase the HA operating system software offerings from the operating system vendor. The following restrictions apply to clusters with HA admin nodes:

- The cluster manager supports RHEL and SLES, but not CentOS, on HA admin nodes.
- The cluster manager requires HA admin nodes to be x86_64 HPE ProLiant DL360 servers.

A cluster can have leader nodes, and these leader nodes can be either scalable unit (SU) leader nodes or ICE leader nodes. ICE leader nodes can be configured as HA leader nodes. The following information applies to clusters with leader nodes:

- The operating system used on the leader nodes must match the operating system used on the admin node.
- The cluster manager supports only x86_64 servers as scalable unit (SU) leader nodes.
- The cluster manager supports RHEL and SLES, but not CentOS, on HA ICE leader nodes (rack leader controllers).

For more information about operating system interoperability, DVDs, and `.iso` file names, see the HPE Performance Cluster Manager release notes.

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

Reinstalling the major cluster system software components quickly

There are situations in which you might want to reinstall all or most of the software on a factory-configured cluster. For example, use this procedure if the following are true:

- You are satisfied with the cluster configuration.

  and

- You want to upgrade the cluster operating system to the next major release.

The procedure in this topic explains how to back up and restore the following:

- The current operating system
- The cluster manager
- The slots
- Other software from a factory-installed cluster system

The procedure in this topic assumes the following:

- You want to reinstall the cluster system as it was configured at the factory with a minimum of changes.
- The cluster is intact, and you can back up the configuration files you need.

If you cannot use this quick reinstallation procedure, proceed to the following topic:

Installing the operating system and the cluster manager jointly

Procedure

1. Use the following command to back up the cluster definition file:

   ```
   # discover --show-configfile [--images] [--kernel] --bmc-info \
   [--kernel-parameters] [--ips] > cdf_backup_location
   ```

   Specify the `--images`, `--kernel`, and `--kernel-parameters` if you plan to reinstall the same operating system images and kernel parameters.

   Specify `--ips` if you want to retain the IP addresses currently assigned. If you omit the `--ips` parameters, the installer allocates an IP address for each node when the `discover` command runs. You can change these IP addresses later with the `cm node modify` command.

   For *cdf_backup_location*, specify a file name.

   For example:

   ```
   # discover --show-configfile --images --kernel --bmc-info \
   --kernel-parameters --ips > my.config.file
   ```

2. Move the cluster definition file backup file to another server at your site.

3. 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. The command writes the backup file to the current directory.

For example:

```
# sqlite3 /opt/clmgr/database/db/cmu.sqlite3 ".backup cmu.backup.sqlite3"
```

4. Move the cluster database backup file to another server at your site.

5. (Conditional) Reset the management switches.

   Complete this step if any of the following conditions exist:

   - You updated or changed the cabling throughout the cluster.
   - You want different or new VLAN numbering to be used across the cluster.
   - You want to update or adjust the IP address subnet ranges used by components throughout the cluster.

   The substeps are as follows:

   a. Back up the switch configuration information that currently exists:

      ```
      # switchconfig config --pull -s all
      ```

      If you need to restore settings on the switches, this step ensures that you have a backup. The command in this step writes the switch configuration files to the following default location on the admin node:

      ```
      /opt/clmgr/tftpboot/mgmtsw_config_files/mgmtswX/config_file
      ```

   b. Remove any redundant cabling between the following:

      - Between management switches. That is, between one switch and another switch.
      - Between management switches and the chassis management controllers (CMCs).

      NOTE: This step addresses clusters with redundant cabling. In these clusters, two management switches are connected by using two or more cables in a bonded cabling pair.

      If the cluster has redundant cabling and you want to reset the management switches to factory settings, it is likely that a networking loop will exist after you reset the switches. This loop causes the network to be unusable until it is resolved. Before you run a factory reset on all management switches, physically remove any redundant cabling between all switches. You can restore the cabling after you rerun the discover command to reconfigure.

   c. Reset the management switches to factory defaults. Start with the highest-numbered management switches and move backwards.

      For example, if a cluster had four management switches named mgmtsw0, mgmtsw1, mgmtsw2, and mgmtsw3, the commands are as follows:

      ```
      # switchconfig reset_factory_defaults -s mgmtsw3 --force
       # switchconfig reset_factory_defaults -s mgmtsw2 --force
       # switchconfig reset_factory_defaults -s mgmtsw1 --force
       # switchconfig reset_factory_defaults -s mgmtsw0 --force
      ```

   d. Wait about 3-10 minutes for the management switches to become reachable again.

      Enter the following command to reach an individual switch:

      ```
      ping mgmtswX
      ```

6. Move the switch configuration information backup file to another server at your site.

7. Enter the following command to display the repositories:

   ```
   # cm repo show
   ```

   The command returns the names of all software repositories on the system.

8. Back up the system software repositories that you need to another system at your site.

   By preserving these repositories, you avoid having to download software from the support websites of HPE and other software distributors.

9. Install the cluster software on the admin node.

   a. Complete the procedures in the following topics:

      - **Installing the operating system and the cluster manager jointly**
      - **Configuring a high availability (HA) admin node**. Complete these procedures if the admin node is an HA admin node.
      - **Installing the cluster software on the admin node**

   b. (Conditional) Add operating system updates or cluster manager updates.

10. Preserve the existing switch configuration.

    Complete this step if you did not complete Step **5**. That is, complete this step if you want to retain the current network, VLAN, and IP configuration in the cluster.

    Enter the following command to direct the `discover` command to omit the switch configuration:

    ```
    # cadmin --enable-discover-skip-switchconfig
    ```

    The command in this step ensures that the `discover` command does not overwrite or configure new settings on the management switches that are added back to the cluster.

11. Run the `discover` command to configure the cluster.

    For information about the `discover` command, see the following:

    Running the `discover` command to complete the cluster configuration

12. (Conditional) Plug in the redundant cables.

    Complete this step if you disconnected the redundant cables earlier in this procedure.

    NOTE: If the network becomes unstable when adding the redundant cabling, you can attempt to reconfigure the switches in the foreground and watch the progress. To watch the progress, enter the following command:

    ```
    switchconfig_configure_node --node mgmtswX
    ```

    For *X*, specify the number of the management switch.

    For example, to reconfigure `mgmtsw0` and `mgmtsw1`, issue the following command:

    ```
    # switchconfig_configure_node --node mgmtsw0,mgmtsw1
    ```

13. (Conditional) Complete the scalable unit (SU) leader cluster configuration.

    Complete this step if this is a cluster with SU leader nodes.

    Complete the procedure in the following:

Completing the scalable unit (SU) leader node configuration

NOTE: Return to this procedure after you complete **Completing the scalable unit (SU) leader node configuration**.

14. Direct the system to enable top-level switch configuration when the `discover` command runs in the future:

    ```
    # cadmin --disable-discover-skip-switchconfig
    ```

15. (Conditional) Recreate custom images or import images that you backed up.

    Complete this step if you have custom repositories that you need to recreate.

    For example, if you have custom repositories for NVIDIA or Mellanox OFED, copy back the repositories that you copied off.

    Import images, add or recreate repositories, and create custom images as necessary.

16. Enter the following command to reboot the cluster:

    ```
    # cm power reboot -t system
    ```

Installing the operating system and the cluster manager separately

Use the deployment procedures in this chapter in the following circumstances:

- You want to install the operating system yourself so you can customize it.
- You use only one slot. This procedure results in only one slot.
- You do not need a high availability admin node.
- You want to install the admin node manually.
- The compute nodes have disks and your intent is to provision those disks.

To start the installation, proceed to the following:

Preparing to install the operating system and the cluster manager separately

Preparing to install the operating system and the cluster manager separately

The following procedure explains how to prepare for the installation.

Procedure

1. Use your hardware documentation to connect the cluster hardware to your site network, and assign roles to each server.

   If you want leader nodes, select one or more nodes to act as leader nodes.

   The admin node needs access to the following:

   - Non-ICE compute nodes
   - Compute node management cards (iLOs) or baseboard management controllers (BMCs)
   - GUI clients

   Although it is not strictly required, each component type typically resides on a separate network. Using independent networks ensures good network performance and isolates problems if network failures occur.

   Configure the NICs on the admin node as follows:

   - Connect one NIC to a network established for compute node administration. The IP address of this NIC is needed during configuration of the admin node.
   - Connect a second NIC to the network connecting the admin node to the GUI clients.
   - A third NIC is typically used to provide access to the network connecting all the compute node management cards (iLOs or BMCs).

2. Download the cluster manager ISO for your operating system, or order a cluster manager media kit from HPE.

The installation instructions assume that you have the cluster manager software on physical media, which can be either a DVD or a bootable USB. If you want to install all your software over a network connection, you do not need to create physical media or to attach a DVD drive. If you install from a network location, modify the instructions accordingly.

If you choose to download the software, use the following instructions to write the software to physical media:

- For a DVD, use your site practices to create the DVD and then attach a DVD drive to the node you want to designate as the admin node.
- For a bootable USB, complete the following steps:

  On a Linux system:

  a. Plug the USB device into the Linux server to which you downloaded the ISO.

  b. In a terminal window, use the following command to retrieve the device name:

  ```
  # dmesg | tail [-20]
  ```

  Specify `-20` on the command if you want the full identity on the USB.

  For example:

  ```
  # dmesg | tail
  [876318.185357] scsi 10:0:0:0: Direct-Access     Lexar    USB Flash Drive  1100 PQ: 0 ANSI: 6
  [876318.185478] scsi 10:0:0:0: alua: supports implicit and explicit TPGS
  [876318.185481] scsi 10:0:0:0: alua: No target port descriptors found
  [876318.185774] sd 10:0:0:0: Attached scsi generic sg5 type 0
  [876318.186994] sd 10:0:0:0: [sdd] 31285248 512-byte logical blocks: (16.0 GB/14.9 GiB)
  [876318.187603] sd 10:0:0:0: [sdd] Write Protect is off
  [876318.187609] sd 10:0:0:0: [sdd] Mode Sense: 43 00 00 00
  [876318.188181] sd 10:0:0:0: [sdd] Write cache: enabled, read cache: enabled, doesn't support DPO or FUA
  [876318.198875] sdd: sdd1 sdd2 sdd3
  [876318.201520] sd 10:0:0:0: [sdd] Attached SCSI removable disk
  ```

  In the preceding example, the device name is `sdd`.

  c. Enter the following commands to find the `/dev/sd`*`X`* of the USB device:

  ```
  # dd if=/dev/zero /dev/sdX bs=512 count=65536
  # dd if=cm-admin-install-1.3.1-os.iso of=/dev/sdX bs=1024
  ```

  For *os*, specify the operating system.

  d. Extract the USB device and plug it in again.

  e. Enter the `parted` command as shown in the following example, and at the `parted` prompt, enter `p` to print the partition map:

  ```
  # parted /dev/sdX
  GNU Parted 3.2 Using /dev/sdd Welcome to GNU Parted! Type 'help' to view a list of commands.
  (parted) p
  ```

  f. (Conditional) Enter `F` to fix the error if there is an error notification.

  If the following message appears, enter `F` to fix:

  ```
  Warning: Not all of the space available to /dev/sdd appears to be used, you can fix the
  GPT to use all of the space (an extra 17098052 blocks) or continue with the current setting?
  Fix/Ignore? F
  ```

  g. Enter `q` to quit.

  On a Windows system, the following procedure uses Win32DiskImager:

a. Plug the USB device into the Windows system to which you downloaded the ISO.

b. Start Win32DiskImager.

c. Click the file folder icon.

d. In the **Select a disk image** popup, browse to the `.iso` file, select the `.iso` file, and click **Open**.

e. In the **Image File** field, verify the path to the location of the `.iso` file.

f. In the **Device** field, verify the destination device.

g. Click **Write**.

NOTE: If a popup window prompts you to format the disk, select **Cancel**. This window can appear multiple times.

h. When the **Complete** popup appears, click **OK**.

3. Plug the USB device into the admin node or mount the DVD.
4. Obtain the operating system installation software.

   For information about the operating system installation software, see the following:

   Operating system releases supported in the HPE Performance Cluster Manager 1.3.1 release
5. Obtain the cluster manager installation software from the following website:

   https://www.hpe.com/downloads/software
6. Proceed to the following:

   Installing and configuring the operating system

Installing and configuring the operating system

The following procedure explains how to install an operating system on the admin node. The procedure notes customizations that the cluster manager requires.

Procedure

1. Install an operating system on the admin node with the following characteristics:

   - Create a static IP address on the admin node.
   - Configure the admin node to use the network time protocol (NTP) server at your site. Configure the time zone for your site.
   - Set the admin node to use your site domain name server (DNS).
   - For the internal traffic between the admin node and other nodes, allow all incoming and outgoing traffic. Configure the admin node NIC as a trusted interface or internal zone.
   - If you install the RHEL operating system on the admin node, do not configure SELinux. The cluster manager disables SELinux.
   - Configure the root file system with enough space to hold all the system images the cluster needs.

- Design the operating system as a conventional operating system with typical installation packages.
- Ensure that only Java version 1.8.0 packages are selected and installed.

2. Proceed to the following:

   Installing the cluster manager

Installing the cluster manager

The following procedure explains how to run the installation script that installs the cluster manager on the admin node.

Procedure

1. Mount the cluster manager admin installation DVD (physical media) or `.iso` image (electronic software).

   Select the appropriate DVD from the HPE Performance Cluster Manager media kit for your target operating system and architecture. Insert the DVD into a DVD reader attached to the admin node, and mount the DVD.

   Alternatively, download the product electronically as an `.iso` file. The `.iso` files on the HPE website use the HPE part number format (for example, `Q9V62-11049.iso`). You can rename the files before or after download. If you download the software as an `.iso` file, use the `mount` command to mount the files and give the download a more descriptive name.

   In the following example, the `mount` command specifies a new, more descriptive name for the `.iso` files:

   ```
   # ls -lh
   .
   .
   -rw-r---r-- 1 linuxdev linuxdev 6.5G Apr 1 02:47 cm-admin-install-1.3.1-rhel8-x86_64.iso
   .
   .
   .
   # mount -o ro,loop cm-admin-install-1.3.1-rhel8-x86_64.iso /mnt
   ```

   For more information about HPE part numbers, see the cluster manager release notes.

2. Enter the following command to change your working directory to the mount point:

   ```
   # cd /mnt
   ```

3. Enter the following command to start the installation script:

   ```
   # ./standalone-install.sh
   ```

   The installation script starts and begins the installation. The script is included on the cluster manager admin installation DVDs and corresponding `.iso` files. The script prompts you for information from time to time. Respond to the prompts with information about your cluster environment.

4. Reboot the cluster.

   For example:

   ```
   # reboot
   ```

5. Proceed to the following:

   Installing the cluster software on the admin node

Installing the operating system and the cluster manager jointly

Use the deployment procedures in this chapter in the following circumstances:

- You received new, cabled hardware from HPE, but no software is installed on the cluster.
- You want to configure custom partitions on the admin node, and you want the installer to configure the operating system and cluster manager together. This method assumes that you want to use the standard operating system installation parameters that are defined in the cluster manager software.
- You want to configure a highly available admin node.
- You want to configure two or more slots.
- A disaster occurred at your site, and you need to recover your cluster.

To start the installation, proceed to the following:

Preparing to install the operating system and the cluster manager jointly

Preparing to install the operating system and the cluster manager jointly

The following procedure explains how to prepare for the installation.

Procedure

1. Obtain the following information for each node:

   - The current username and password for each **management card**. The management card is either a baseboard management controller (BMC) or an iLO component.
   - The MAC address of the management card of each node.
   - The MAC address of the node.

   You can obtain this information in one of the following ways:

   - From the HPE factory. The factory has a list of MAC addresses. In addition, you might have ordered the cluster with a fixed username and password for each node.
   - From the node itself. Apply power to the management card in the node, access the management card, and read the MAC address from the BIOS menu.
   - From the sticker attached to the node. The sticker bears the MAC address.

2. Contact your site network administrator to obtain network information for the management card in the admin node.

   For the admin node management card, obtain the following:

- (Optional) The current IP address of the management card on the admin node. If you do not have this information, you can set the management card address from a serial console.
- The IP address you want to set for the management card.
- The netmask you want to set for the management card.
- The default gateway you want to set for the management card.
- A hostname.
- The domain name.
- An IP address.
- The netmask.
- The default route.
- The root password.

Obtain the following information about your site network:

IP addresses of the domain name servers (DNSs)

NOTE: To configure two nodes, as part of a two-node HA admin node configuration, make sure to obtain the necessary configuration information for both nodes.

3. Attach a DVD drive to the admin node.

The installation instructions assume that you have the cluster manager software on physical media. You can order a cluster manager media kit from HPE. Alternatively, you can download the cluster manager ISO for your operating system and use your site practices to create a DVD from the ISO.

If you want to install all your software over a network connection, you do not need to attach a DVD drive. If you install from a network location, modify the instructions accordingly.

4. Obtain the operating system installation software.

For information about the operating system installation software, see the following:

Operating system releases supported in the HPE Performance Cluster Manager 1.3.1 release

5. Obtain the cluster manager installation software from the following website:

https://www.hpe.com/downloads/software

The website requires you to log in with your HPE Passport account.

6. (Conditional) Configure the storage unit hardware and software.

Complete this step only if you want an HA admin node.

The HA admin node environment requires an HPE MSA 2050 storage unit and associated software.

When you configure the storage unit, configure one LUN per slot. If your cluster was configured at the HPE factory, the factory configured the storage unit for one LUN per slot.

You can manage the storage unit from one of the physical admin nodes or from another computer. For example, you can use a laptop to manage the storage unit. The following explain other management methods:

- You can install the storage unit software on one or both of the physical admin nodes. In this case, you can manage the storage unit in-band or out-of-band from one of the physical admin nodes.

To manage the storage unit in-band, connect the storage controllers to the management switches.

To manage the storage unit out-of-band, attach the storage unit to the network, and address the controllers on the site network.

- You can install the storage unit software on a computer outside the cluster. In this case, you can manage the storage unit out-of-band from either of the physical admin nodes. To manage, address the controllers on the public network.

After you install the storage unit software, start the storage unit software GUI to add the addresses and passwords of the storage controllers.

7. Attach the cluster to your site network.

   Use the procedure in the following manual:

   HPE Performance Cluster Manager Getting Started Guide

8. If one is available, retrieve the cluster definition file for this cluster.

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

   The cluster definition file can reside in any directory, under any name, on the cluster. By default, the HPE factory installation writes the cluster definition file to the following location:

   ```
   /var/tmp/mfgconfigfile
   ```

   If the configuration file is no longer in that location, use the following command to create a cluster definition file and write it to a location of your own choosing:

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

   For *filename*, specify the output file name. This command writes the cluster definition file to *filename*.

   If you backed up the cluster definition file, use the backup copy at your site. If necessary, you can obtain a copy of the original cluster definition file from the HPE factory.

9. Proceed to one of the following:

   - To configure custom partitions on the admin node, proceed to the following:

     (Conditional) Configuring custom partitions on the admin node

   - To install the admin node with default partitioning, proceed to the following:

     Inserting the installation DVD and booting the system

(Conditional) Configuring custom partitions on the admin node

Complete the procedure in this topic if the default partitioning scheme does not suit the needs of this cluster.

If you create custom partitions on the admin node, you can create custom partitions on one or more non-ICE compute nodes. The partitions on the non-ICE compute nodes can be different from the partitions on the admin node. If you accept default partitions on the admin node, you can still create custom partitions on the non-ICE compute nodes.

The procedure in this topic explains how to specify custom partitions for the admin node. When the admin node boots, the boot process creates the partitions. A later procedure explains how to run the `discover` command to configure the nodes. When you run the `discover` command, you can create the same (or different) custom partitions on the non-ICE compute nodes. While you can create custom partitions on leader nodes, Hewlett Packard Enterprise recommends that you accept the default partitions on leader nodes. Because ICE compute nodes are diskless, they cannot be partitioned.

NOTE: If you choose to implement custom partitions on the admin node, the admin node is reduced to one slot. Keep this caveat in mind if you want to configure custom partitions on the admin node. HPE does not support custom admin node partitions on clusters with HA admin nodes or HA ICE leader nodes. For information about the default cluster partitioning scheme, see the following:

Default partition layout information

The following procedure explains how to create custom partitions on the admin node.

Procedure

1. Mount the cluster manager installation DVD into the DVD drive of a local computer at your site.

   Do not mount the installation DVD into the DVD drive on the cluster.

2. Read all the information in `README.install` file.

   This file resides in the root directory of the installation DVD.

   This file includes general installation and custom partitioning information.

3. Read all the information in `custom_partitions_example.cfg`.

   This file resides in the root directory of the installation DVD.

   This file contains information about how to use the file and about the effect of custom partitions on cluster operations.

   When you install an admin node with custom partitions, the installer destroys all other data. The destroyed data includes any slot specifications that might reside on the admin node hard disk. In other words, when you install an admin node with custom partitions, you no longer have a cluster with slots. By extension, when the admin node is configured with custom partitions, you cannot have non-ICE compute nodes with multiple slots.

4. Decide where you want the custom configuration file to reside.

   The file can reside on an NFS server or on the installation media. Write `custom_partitions_example.cfg` as follows:

   - To write the configuration file to an NFS server at your site, use an existing server. A later procedure explains how to specify the location to the installer at boot time.
   - To write the configuration file to the installation media, contact your HPE representative for instructions.

5. Open file `custom_partitions_example.cfg` in a text editor, and specify the partitions you want for the admin node.

   The `custom_partitions_example.cfg` file consists of columns of data separated by vertical bar (`|`) characters, which separate the fields into columns. Be careful with the columns in this file. All vertical bar characters must align in order for the partitioning to complete correctly.

   For the `/var` partition, make sure to specify enough size to create and host the images you need for the nodes.

   The file system specifications that the cluster manager supports are as follows:

   - `XFS`
   - `ext4`, which is the default root file system for the cluster manager
   - `ext3`

6. Save and close the file as `custom_partitions.cfg`

7. Proceed to the following:

Inserting the installation DVD and booting the system

You can configure the cluster to boot from up to 10 slots. A slot consists of all the partitions related to a Linux installation.

On a factory-configured cluster, the default number of slots is as follows:

- The default number of slots is two on a cluster with leader nodes.
- The default number of slots is one on a cluster without leader nodes.

Multiple slots, especially on the admin node, can lead to a smoother update when it is time to upgrade the cluster manager or operating system software.

When the cluster is configured with two or more slots, you can clone a production slot to an alternative location, thus creating a fallback slot.

The following topics explain how to select the correct number of slots and the correct boot options for your cluster:

- **About slots**.
- **Booting the system**. Complete the procedure in this topic after you have determined the number of slots to create and the boot options you need.

About slots

A multiple-slot disk layout creates the same disk layout on all nodes. Each slot includes the following:

- A `/boot` partition.
- A `/`, or root, partition.
- A `/boot/efi` partition. A slot includes this partition only if the node is an EFI node.

When you insert the cluster manager operating system installation disk and power on the admin node, you can select a boot method from the GNU GRUB menu. If you select **Install: Wipe Out and Start Over: Prompted**, the installer creates two slots and writes the initial installation to slot 1. After the system is installed, you cannot change the number of slots. If you attempt to change the number of slots, you destroy the data on the disks.

After you install a multislot cluster, you can boot the cluster with the operating system of your choice. This capability might be useful if you ever want to test an operating system or other software. When you have more than one slot, you can roll back an upgrade completely.

The following are some other characteristics of multiple-slot systems and single-slot systems:

Multiple-slot	Single-slot
You can install different operating systems, or different operating system versions, into different slots. If you have leader nodes, the admin node and the leader nodes must have the same operating system installed.	You can install only one operating system for the entire cluster.
As you increase the number of slots, you decrease the amount of disk space per slot. Hewlett Packard Enterprise recommends a minimum of 100 GB per slot.	A single slot uses all available disk space.

Booting the system

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

Procedure

1. Ensure that the admin node is configured to boot from a DVD.

   If necessary, attach an external DVD reader to the admin node.

2. Insert the cluster manager installation DVD into the DVD drive attached to the admin node.

   This DVD has an operating-system-specific label.

   For information about the operating system installation software, see the following:

   Operating system releases supported in the HPE Performance Cluster Manager 1.3.1 release

3. Power on the admin node.

4. Use the arrow keys to select **Display Instructions**, and read the instructions carefully.

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

   Each boot option has a set of default behaviors. Some boot options permit you to specify custom boot parameters. The options are as follows:

   - **Display Instructions**

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

   - **Install: Install to Designated Slot**

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

     This boot option permits you to specify custom boot parameters.

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

     Select this option if you want to add slots.

     This option destroys all information currently on the cluster. The installer partitions the admin node with the specified number of slots, and the installer writes the initial installation to the designated slot. For example, for an initial installation, select this option.

   - **Rescue: Prompted**

To create a troubleshooting environment, select this option.

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

  Select this option if you want to customize the installation. This option lets you supply all boot options as command-line parameters. Unlike the other boot methods, there are no system prompts for boot options. More information is available in **Display Instructions**.

  This boot option permits you to specify custom boot parameters. Hewlett Packard Enterprise recommends this option only for users with installation experience.

Example 1. To specify `console=` or any other custom boot parameter, select the **Display Instructions** option. Familiarize yourself with the parameters you want to use before you select an actionable option.

Example 2. To allocate scratch disk space on the system disk of the admin node, add the following parameters to the kernel parameter list:

- `destroy_disk_label=yes`
- `root_disk_reserve=`*`size`*

For *size*, specify a size in GiB. The cluster manager creates the scratch disk space in partition 61, but you must otherwise structure the scratch disk space. That is, you create the file system, add the `fstab` entries, and so on. For more information about how to create scratch disk space for a node, see the following:

HPE Performance Cluster Manager Administration Guide

Example 3. To configure this node as one of the physical nodes in an HA admin node, select **Install: Wipe Out and Start Over: Prompted**.

6. Respond to the questions that the installation menus ask.

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

   - **Enter number of slots to allow space for: (1-10):**

     Enter `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`, or `10`.

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

     Typically, you want at least two slots.

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

     Enter `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`, or `10`.

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

     If you selected **Install: Wipe Out and Start Over: Prompted**, you can select slot 1.

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

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

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

   - **Is this a physical admin node in an SAC-HA configuration? (normally no) (y/n):**

     To configure this node as part of an HA admin node configuration, enter **`y`** and press Enter.

If this node is a standalone, non-HA admin node, enter **n** and press Enter.

- **Use predictable network names for the admin node? (normally yes) (y/n):**

  This dialog question determines whether predictable names or legacy names are assigned to the network interface cards (NICs) in the node.

  To configure the admin node with predictable names, enter **y** and press Enter. The following types of nodes can use predictable names:

  - Standalone admin nodes
  - The virtual machine admin node that is part of a high availability (HA) admin node

  Hewlett Packard Enterprise recommends that you enter **y** when possible.

  To configure a physical admin node as part of a two-node HA admin node, enter **n** and press Enter. This action configures the node with legacy names. A physical admin node that is part of an HA admin node requires legacy names.

  For information about predictable network names, see the following:
  Predictable network interface card (NIC) names

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

  To specify additional boot parameters, enter them in a comma-separated list and press Enter.

  - Example 1. To configure custom partitions, add the target for the custom partitions file, as follows:

    `custom_partitions=`*`NFS_server_address`*`:/`*`path_to_custom_partitions.cfg`*

Variable	Specification
NFS_server_address	The identifier of your site NFS server. This address can be an IP address or a hostname.
path_to_custom_partitions.cfg	The full path to the custom partitioning file.

  - Example 2. For an HA admin node based on an HPE ProLiant DL360, specify the following:

    `console=ttyS0,115200n8`

  This documentation does not describe a network install of an HA admin node. However, if you install an HA admin node over the network, specify `sac_ha=1` as a boot parameter.

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

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

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

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

7. Wait for the installation to complete.

   The installation can take several minutes.

8. Remove the operating system installation DVD.

9. At the # prompt, enter **reboot**.

This boot is the first boot from the admin node hard disk.

10. Proceed to one of the following:

    - To configure the RHEL operating system, proceed to the following:

      Configuring RHEL 8.X or RHEL 7.X on the admin node

    - To configure the SLES operating system, proceed to the following:

      Configuring SLES 15 SPX and SLES 12 SPX on the admin node

Configuring RHEL 8.X or RHEL 7.X on the admin node

The following procedure explains how to configure RHEL 8.X or RHEL 7.X on the admin node.

Procedure

1. At the **Username** prompt, enter `root`.
2. At the **Password** prompt, enter `cmdefault` or whatever password you set.

   The default password is `cmdefault`.
3. Open a VGA window to the admin node.
4. On the **Welcome!** screen, complete the following steps:

   - Select your language.
   - Click **Next**.
5. On the **Typing** screen, complete the following steps:

   - Select your keyboard.
   - Click **Next**.
6. On the **Privacy** screen, click **Next**.
7. On the **Connect Your Online Accounts** screen, click **Skip**.
8. On the **Getting Started** screen, click **X**.

   This action dismisses the screen.
9. On the **You're ready to go!** screen, click **Start using Red Hat Enterprise Linux Server**.
10. Open a terminal window or `ssh` session to the admin node.
11. Use a text editor to open file `/etc/hosts`.
12. Add a line that contains address information for this node.

    In file `/etc/hosts`, format the line as follows:

    admin_node_IP admin_node_FQDN admin_node_hostname

    The variables are as follows:

Variable	Specification
admin_node_IP	The IP address of the admin node
admin_node_FQDN	The fully qualified domain name (FQDN) of the admin node
admin_node_hostname	The hostname of the admin node

For example, assume that this node is your only physical admin node. Or, assume that this node is the first physical node in a two-node HA admin node configuration. Add the following line:

```
100.162.244.88 acme-admin.acme.usa.com acme-admin
```

13. (Conditional) Add another line of address information for the second physical HA admin node.

    Complete this step if you have a second physical admin node. In this case, you want to configure the second physical node in an HA admin node configuration.

    Example 1. In file `/etc/hosts`, add the following line:

    ```
    100.162.244.89 acme-admin2.acme.usa.com acme-admin2
    ```

    Example 2. You can add a comment line to precede the node identification lines. For example, your new lines might look as follows:

    ```
    #physical node addresses
    100.162.244.88 acme-admin1.acme.usa.com acme-admin
    100.162.244.89 acme-admin2.acme.usa.com acme-admin2
    ```

14. (Conditional) Add identifying lines for the storage unit.

    Complete this step if you are configuring two physical admin nodes into an HA admin node.

    Add two lines in the following format:

    ```
    is_1_IP is_1_FQDN is_1_hostname
    is_2_IP is_2_FQDN is_2_hostname
    ```

    The variables are as follows:

Variable	Specification
is_1_IP	The IP address of the first storage unit
is_1_FQDN	The fully qualified domain name (FQDN) of the first storage unit
is_1_hostname	The hostname of the first storage unit
is_2_IP	The IP address of the second storage unit
is_2_FQDN	The fully qualified domain name (FQDN) of the second storage unit
is_2_hostname	The hostname of the second storage unit

For example, the following lines describe both admin nodes and the storage unit:

```
#physical node addresses
100.162.244.88 acme-admin1.acme.usa.com acme-admin
100.162.244.89 acme-admin2.acme.usa.com acme-admin2
#
# IS consoles
#
100.166.33.138  toki-stor-1.acme.com  toki-stor-1   # IS console
100.166.33.139  toki-stor-2.acme.com  toki-stor-2   # IS console
```

15. Save and close file `/etc/hosts`.

16. Enter the following command to set the admin node hostname:

```
# hostnamectl set-hostname admin_node_hostname
```

For *admin_node_hostname*, make sure to enter the hostname, which is the short name. Do not enter the FQDN, which is the longer name.

If you complete this step as part of a two-node HA admin node configuration, specify the *admin_node_hostname* of the node you are configuring at this time.

17. Enter the following command to create file `/etc/sysconfig/network` with no content:

```
# touch /etc/sysconfig/network
```

18. Use the `ip addr show` command to determine the following:

- The name of the network interface card (NIC) that connects the admin node to the house network.
- The MAC address of the NIC that connects the admin node to the house network.

For example, in the following output, the NIC name is `ens20f0` and the MAC address is `00:25:90:fd:3d:a8`:

```
admin # ip addr show
1: lo:  mtu 65536 qdisc noqueue state UNKNOWN qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: ens20f0:  mtu 1500 qdisc mq state UP qlen 1000
    link/ether 00:25:90:fd:3d:a8 brd ff:ff:ff:ff:ff:ff
    inet 128.162.243.106/24 brd 128.162.243.255 scope global ens20f0
       valid_lft forever preferred_lft forever
    inet6 fe80::225:90ff:fefd:3da8/64 scope link
       valid_lft forever preferred_lft forever
3: ens20f1:  mtu 1500 qdisc mq master bond0 state UP qlen 1000
    link/ether 00:25:90:fd:3d:a9 brd ff:ff:ff:ff:ff:ff
4: ens20f2:  mtu 1500 qdisc mq master bond0 state DOWN qlen 1000
    link/ether 00:25:90:fd:3d:a9 brd ff:ff:ff:ff:ff:ff
5: ens20f3:  mtu 1500 qdisc mq state DOWN qlen 1000
    link/ether 00:25:90:fd:3d:ab brd ff:ff:ff:ff:ff:ff
.
.
.
```

19. Use a text editor to open file `ifcfg-`*`name`*.

File `ifcfg-name` is the configuration file for the NIC that is connected to the house network. For example, `ifcfg-ens20f0`.

The path to this file is as follows:

```
/etc/sysconfig/network-scripts/ifcfg-name
```

20. In the `ifcfg-name` file, update the following lines with the information for this cluster:

```
NAME=name      # Add the name of the house NIC
DEVICE=name    # Add the name of the house NIC
IPADDR=        # Add the IP address of this admin node
PREFIX=        # Add your site netmask setting.  For example:  24
GATEWAY=       # Add your site gateway
DNS1=          # Add your site primary DNS IP address
DNS2=          # Add your site secondary DNS IP address
DOMAIN=        # Add your site domain
HWADDR=        # Add the NIC MAC address
BOOTPROTO=     # Set to "none"
ONBOOT=        # Set to "yes"
DEFROUTE=      # Set to "yes"
TYPE=          # Set to "Ethernet"
UUID=          # Enter "nmcli connection show" to retrieve the UUID value
```

For information about how the inputs to the `ifcfg-name` file, see the following:

- Your RHEL documentation
- The `nm-settings-ifcfg-rh` manpage

NOTE: The cluster software does not support IPV6 on the public NIC in the admin node. The following line is needed for this installation:

```
IPV6INIT="no"
```

You can remove the lines that start with `IPV6_` from the `ifcfg-name` file, or you can retain those lines for completeness.

21. Save and close file `ifcfg-name`.

22. Bring up the NIC with the updated networking information.

Use `ifdown` and `ifup` commands in the following format:

```
ifdown name
ifup name
```

For example:

```
# ifdown ens20f0
# ifup ens20f0
```

23. Enter the following command to restart the name service cache daemon, `ncsd`:

```
# systemctl restart nscd
```

24. Enter the following command to retrieve current time zone information:

```
# date
Fri Apr 20 10:12:50 CDT 2019
```

The previous output is an example that shows the admin node set to US central daylight time. If the output you see is not correct for this cluster, complete the following steps:

- Enter the following command to display a list of time zones:

  ```
  # timedatectl list-timezones
  ```

- Use the following command to set the time zone:

  ```
  timedatectl set-timezone time_zone
  ```

  For *time_zone*, specify one of the time zones from the `timedatectl list-timezones` command output.

  When finished, you can use the `timedatectl` command to display the time zone information you configured. For example:

  ```
  # timedatectl
  Local time: Fri 2019-04-15 14:55:33 PDT
  Universal time: Fri 2019-04-15 21:55:33 UTC
  RTC time: Fri 2019-04-15 21:55:33
  Time zone: America/Los_Angeles (PDT, -0700)
  NTP enabled: yes
  NTP synchronized: yes
  RTC in local TZ: no
  DST active: yes
  Last DST change: DST began at
  Sun 2019-03-13 01:59:59 PST
  Sun 2019-03-13 03:00:00 PDT
  Next DST change: DST ends (the clock jumps one hour backwards) at
  Sun 2019-11-06 01:59:59 PDT
  Sun 2019-11-06 01:00:00 PST
  ```

- Enter the following command to confirm the time zone:

  ```
  # date
  Fri Apr 20 10:12:50 PDT 2019
  ```

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

The following steps direct requests to your site NTP server instead of to the public time servers of the `pool.ntp.org` project:

- Use a text editor to open file `/etc/chrony.conf`.
- Insert a pound character (#) into column 1 of each line that includes `rhel.pool.ntp.org`.

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

- At the end of the file, add a line that points to your site NTP server.

  The following is an example of a correctly edited file:

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

```
# server 2.rhel.pool.ntp.org
server ntp.mycompany.com iburst
```

- Enter the following line to allow clients on the management network to query the NTP server:

  ```
  allow 172.23
  ```

- Enter the following command to restart the NTP server:

  ```
  # systemctl restart chronyd
  ```

26. Proceed to one of the following:

- If you have only one admin node, proceed to the following:

  Specifying the cluster configuration

- If you have two admin nodes, then you are configuring this node as part of a two-node HA admin node. If this node is the first of the two nodes, log into the second physical admin node, and install the operating system software on the other node. Proceed to the following:

  Inserting the installation DVD and booting the system

- If you have two admin nodes, then you are configuring this node as part of a two-node HA admin node. If this node is the second of the two nodes, proceed to the following:

  Configuring a high availability (HA) admin node

Configuring SLES 15 SPX and SLES 12 SPX on the admin node

The procedure in this topic uses the SLES YaST interface. To navigate YaST, use key combinations such as the following:

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

For more information about navigation, see the following:

YaST navigation

Procedure

1. Connect to the admin node.

Complete this step if you are not already logged into the admin node.

Use one of the following methods:

- Move to the room in which the cluster resides and log into the admin node
- Through the intelligent platform management interface (IPMI) tool

- Through the console attached to the cluster
- Through a separate keyboard, video display terminal, and mouse

2. Start YaST2.

   Enter the following commands:

   ```
   # export Textmode=1
   # export TERM=xterm
   # /usr/lib/YaST2/startup/YaST2.Firstboot
   ```

   In addition, you might need to alter your environment. For example to run YaST from a PuTTY window, enter the following:

   ```
   # export NCURSES_NO_UTF8_ACS=1
   ```

3. On the **Language and Keyboard Layout** screen, complete the following steps:

   - Select your language.
   - Select your keyboard layout.
   - Select **Next**.

4. On the **Welcome** screen, select **Next**.

5. On the **License Agreement** screen for the operating system, complete the following steps:

   - Tab to the box (`[ ] I Agree ...`).
   - Press the spacebar to accept the license terms. This action puts an `x` in the box, so it looks like this: `[x]`.
   - Select **Next**.
   - (Conditional) If there are more license agreement screens, select **Next** again.

6. On the **Network Settings** screen, specify the NIC information.complete the following steps:

   This step differs, depending on the operating system, as follows:

   - For SLES 15 SPX systems, complete the following steps:

     - Highlight the NIC with the lowest MAC address. Look at the final octet in each MAC address.

       For example, if the node includes the following NICs, highlight the NIC numbered `ec:eb:b8:89:f2:90`:

       ```
       hikari2:~ # ip addr | grep ether
           link/ether ec:eb:b8:89:f2:90 brd ff:ff:ff:ff:ff:ff    # lowest
           link/ether ec:eb:b8:89:f2:91 brd ff:ff:ff:ff:ff:ff
           link/ether ec:eb:b8:89:f2:92 brd ff:ff:ff:ff:ff:ff
           link/ether ec:eb:b8:89:f2:93 brd ff:ff:ff:ff:ff:ff
       ```

     - Select **Edit**.

   - For SLES 12 SPX systems, complete the following steps:

    - Highlight the first NIC that appears underneath **Name**.
    - Select **Edit**.

7. On the **Network Card Setup** screen, specify the admin node public NIC.

   Complete the following steps:

   - Select **Statically Assigned IP Address**. Hewlett Packard Enterprise recommends a static IP address, not DHCP, for the admin node.
   - In the **IP Address** field, enter the admin node IP address. This IP address is the IP address for users to use when they want to access the cluster.
   - In the **Subnet Mask** field, enter the admin node subnet mask.
   - In the **Hostname** field, enter the admin node fully qualified domain name (FQDN). HPE requires you to enter an FQDN, not the shorter hostname, into this field. For example, enter `mysystem-admin.mydomainname.com`. Failure to supply an FQDN in this field causes the `configure-cluster` command to fail.
   - Select **Next**.

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

8. On the **Network Settings** screen, complete the following steps:

   - Select **Hostname/DNS**.
   - In the **Hostname** field, enter the admin node hostname.
   - (SLES 12 SPX only) In the **Domain Name** field, enter the domain name for your site.
   - (SLES 12 SPX only) Put an `X` in the box next to **Assign Hostname to Loopback IP**.
   - In the **Name Servers and Domain Search List**, enter the IP addresses of the name servers for your house network.
   - In the **Domain Search** field, enter the domains for your site.
   - Back at the top of the screen, select **Routing**.

     The **Network Settings > Routing** screen appears.
   - In the **Default IPV 4 Gateway** field, enter your site default gateway.
   - Select **Next**.

9. On the **Clock and Time Zone** screen, complete the following steps:

   - Select your region.
   - Select your time zone.
   - In the **Hardware Clock Set To** field, choose **Local Time** or accept the default of **UTC**.
   - Select **Next**.

This step synchronizes the time in the BIOS hardware with the time in the operating system. Your choice depends on how the BIOS hardware clock is set. If the clock is set to GMT, the operating system switches between standard time and daylight savings time automatically. GMT corresponds to UTC.

10. On the **Local User** screen, complete one of the following steps:

    - Provide information for additional user accounts and select **Next**.

      or

    - Select **Skip User Creation** and select **Next**.

11. On the **Authentication for the System Administrator "root"** screen (SLES 15 SPX) or on the **Password for System Administrator “root”** (SLES 12) screen, complete the following steps:

    - In the **Password for root User** field, enter the password you want to use for the root user.

      This password becomes the root user password for all the system nodes.

    - In the **Confirm password** field, enter the root user password again.

    - In the **Test Keyboard Layout** field, enter a few characters.

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

    - Select **Next**.

    - (Conditional) Confirm the password on the popup that appears.

      Complete this step if a password popup appears.

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

13. Log into the admin node.

14. Open file `/etc/hosts` within a text editor.

15. Within file `/etc/hosts`, verify that the admin node fully qualified domain name (FQDN) and hostname are correct.

    For example, assume an admin node with the following:

    - An IP address of `100.100.100.100`

    - An FQDN of `mysystem-admin.mydomain.com`

    - A hostname of `mysystem-admin`

    The following `/etc/hosts` file entry describes this admin node correctly:

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

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

16. Confirm that the system is working as expected.

    For example, enter the following command:

    ```
    # ping -c 1 www.hpe.com
    ```

    If necessary, restart YaST to correct settings.

17. (Conditional) Enter a tilde character (~) and then a period character (`.`) to exit the IPMI tool.

Complete this step if your connection is through IPMI.

18. Log into the admin node as the root user.

19. Edit file /etc/chrony.conf to direct requests to the network time protocol (NTP) server at your site.

    The following steps direct requests to your site NTP server instead of to the public time servers of the pool.ntp.org project:

    - Use a text editor to open file /etc/chrony.conf.
    - At the end of the file, add a line that points to your site NTP server.

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

      The following is an example of a correctly edited file:

      ```
      # Use public servers from the pool.ntp.org project.
      # Please consider joining the pool (http://www.pool.ntp.org/join.html).

      server ntp.mycompany.com iburst

      # ! pool pool.ntp.org iburst
      ```

    - Enter the following line to allow clients on the management network to query the NTP server:

      ```
      allow 172.23
      ```

    - Enter the following command to restart the NTP server:

      ```
      # systemctl restart chronyd
      ```

20. (Optional) Configure the system to install the operating system from a VGA screen and perform later operations from a serial console.

    Complete the following steps:

    - Use a text editor to open file /etc/default/grub. By default, the installation DVD specifies certain parameters for GRUB_CMDLINE_LINUX_DEFAULT line and the GRUB_TERMINAL line. The following steps explain what to change on these two lines.
    - On the GRUB_CMDLINE_LINUX_DEFAULT line, edit the line to include console=ttyS0,115200n8 and remove splash=silent quiet. For example:

      ```
      GRUB_CMDLINE_LINUX_DEFAULT="console=ttyS0,115200n8
      intel_idle.max_cstate=1 processor.max_cstate=1 net.ifnames=0
      biosdevname=0 numa_balancing=disable predictable_net_names=0
      intel_iommu=on"
      ```

    - On the GRUB_TERMINAL line, edit the line to change gfxterm to console. For example:

      ```
      GRUB_TERMINAL="console"
      ```

    - Enter the following command to apply the changes made in /etc/default/grub to the GRUB configuration file:

      ```
      # /usr/sbin/grub2-mkconfig -o /boot/grub2/grub.cfg
      ```

    Later, to access the admin node from only a VGA, do the following:

- From the GRUB_CMDLINE_LINUX_DEFAULT= line, remove all the console= parameters.
- Change the GRUB_TERMINAL line back to GRUB_TERMINAL="gfxterm"

21. (Conditional) Open the firewall for port 22 and the ssh command.

Complete this step to install an HA admin node. This step differs depending on your operating system level.

- On SLES 15 SPX systems, enter the following command:

```
# firewall-cmd --zone=external --add-service=ssh
```

- On SLES 12 SPX systems, enter the following command:

```
# yast firewall services add service=service:sshd zone=EXT
```

22. (Optional) Add admin to the No Proxy Domains line (no_proxy= line).

If using a proxy, ensure that admin is added to the No Proxy Domains line in the YaST2 proxy settings for the following:

- The admin node
- The virtual admin nodes of a highly available cluster
- The login nodes

For information about how to configure a proxy, see the SLES documentation.

23. Proceed to the following:

- If you have only one admin node, proceed to the following:

  Specifying the cluster configuration

- If you have two admin nodes, then you are configuring this node as part of a two-node HA admin node. If this node is the first of the two nodes, log into the second physical admin node, and install the operating system software on the other node. Proceed to the following:

  Inserting the installation DVD and booting the system

- If you have two admin nodes, then you are configuring this node as part of a two-node HA admin node. If this node is the second of the two nodes, proceed to the following:

  Configuring a high availability (HA) admin node

Configuring a high availability (HA) admin node

HPE supports your ability to configure HA admin nodes for your cluster. When the cluster is running, the admin node resides in a virtual machine upon one of the physical admin nodes. When a failover occurs, the virtual machine passes from the active node to the passive node.

When you create an HA admin node, you install the cluster manager software, operating system software, and supporting software on two physical admin nodes. After the installation and configuration is complete, the admin node operates within a virtual machine that can reside on either of the two physical hosts.

The general process is as follows:

- Install the operating system software and cluster software on one physical admin node. This phase of the installation process is similar to a non-HA installation, so differences are noted only when necessary.
- Install the operating system software and cluster software on the other physical admin node. Again, this phase of the installation process is similar to a non-HA installation, so differences are noted only when necessary.
- Designate one of the physical admin nodes as the primary admin node, and edit the following file on the primary admin node:

  ```
  /etc/opt/sgi/sac-ha-initial-setup.conf
  ```
- Run the HA admin node initial setup script on the primary physical node.

Complete the following procedures to configure two physical admin nodes to work together as an HA admin node:

- **Configuring the storage unit**
- **Enabling an input-output memory management unit (IOMMU)**
- **Verifying the configuration**
- **Creating and installing the HA software repositories on the physical admin nodes**
- **Preparing to run the HA admin node configuration script**
- **Running the highly available (HA) admin node configuration script**
- **Starting the HA virtual manager and installing the cluster manager on the virtual machine**

NOTE:

- Complete the procedures in this chapter only if you are configuring two admin nodes to work together as an HA admin node.
- You can enable an HA admin node only if both physical admin nodes use the x86_64 architecture. You cannot enable an HA admin node if the physical nodes use the Arm (AArch64) architecture.
- The physical HA admin nodes require legacy network interface card (NIC) names. The virtual machine admin node can use predictable NIC names.

Configuring the storage unit

For the storage unit, the typical configuration is a 2-LUN storage unit. For a cluster with two slots, you can use one LUN per slot.

The following procedure assumes the following about the storage unit:

- The unit is attached to the two admin nodes.
- It is known to be working properly.
- It hosts no content that you want to save. This procedure wipes the storage completely.

Procedure

1. Enter the `lsscsi` command on each physical node to determine the disk devices that each node can recognize.

   In the `lsscsi` output, the storage unit reports as `MSA 2050 SAS`.

   For example, the following output shows that the nodes can recognize disks `/dev/sdc` and `/dev/sdd`. The same disks also appear as `/dev/sde` and `/dev/sdf`, which are secondary paths.

   - On physical node 1, the following output shows that the MSA 2050 devices host `/dev/sdc` and `/dev/sdd`:

     ```
     # lsscsi
     [0:0:0:0]    disk    Generic- SD/MMC CRW       1.00  /dev/sdb
     [15:0:0:0]   enclosu HPE      Smart Adapter    1.04  -
     [15:1:0:0]   disk    HPE      LOGICAL VOLUME   1.04  /dev/sda
     [15:2:0:0]   storage HPE      P408i-a SR Gen10 1.04  -
     [16:0:0:0]   enclosu HP       MSA 2050 SAS     G22x  -
     [16:0:0:1]   disk    HP       MSA 2050 SAS     G22x  /dev/sdc
     [16:0:0:2]   disk    HP       MSA 2050 SAS     G22x  /dev/sdd
     [16:0:1:0]   enclosu HP       MSA 2050 SAS     G22x  -
     [16:0:1:2]   disk    HP       MSA 2050 SAS     G22x  /dev/sde
     [16:0:1:3]   disk    HP       MSA 2050 SAS     G22x  /dev/sdf
     ```

   - On the physical node 2, the following output shows that the MSA 2050 devices host `/dev/sdc` and `/dev/sdd`:

     ```
     # lsscsi
     [0:0:0:0]    disk    Generic- SD/MMC CRW       1.00  /dev/sdb
     [14:0:0:0]   enclosu HPE      Smart Adapter    1.04  -
     [14:1:0:0]   disk    HPE      LOGICAL VOLUME   1.04  /dev/sda
     [14:2:0:0]   storage HPE      P408i-a SR Gen10 1.04  -
     [15:0:0:0]   enclosu HP       MSA 2050 SAS     G22x  -
     [15:0:0:1]   disk    HP       MSA 2050 SAS     G22x  /dev/sdc
     [15:0:0:2]   disk    HP       MSA 2050 SAS     G22x  /dev/sdd
     [15:0:1:0]   enclosu HP       MSA 2050 SAS     G22x  -
     [15:0:1:2]   disk    HP       MSA 2050 SAS     G22x  /dev/sde
     [15:0:1:3]   disk    HP       MSA 2050 SAS     G22x  /dev/sdf
     ```

   The preceding output is an example. The device IDs associated with each disk vary by node and might be different for your configuration.

2. Enter the `pvscan` command on each physical node to determine the disk devices that are initialized and in use currently.

   In the `pvscan` output, the unused devices are **not** listed.

   For example:

- On the first physical node, the following output shows that devices /dev/sdc and /dev/sda2 are in use:

  ```
  # pvscan
  PV /dev/sdc     VG vgha1          lvm2 [100 GiB / 0    free]
  PV /dev/sda2    VG vg_host        lvm2 [200 GiB / 0    free]
  Total: 2 [300 GiB] / in use: 2 [0 GiB] / in no VG: 0 [0   ]
  ```

- On the second physical node the following output shows that devices /dev/sdc and /dev/sde are in use:

  ```
  # pvscan
  PV /dev/sdc     VG vgha1          lvm2 [100 GiB / 0    free]
  PV /dev/sde     VG vg_host        lvm2 [200 GiB / 0    free]
  Total: 2 [300 GiB] / in use: 2 [0 GiB] / in no VG: 0 [0   ]
  ```

3. Based on your analysis of the lsscsi and pvscan commands, choose a disk that both nodes can recognize and that is not currently in use.

A disk that appears in the pvscan output is initialized and might already contain data. Do not select a disk that appears in the pvscan output because it is likely that the disk already contains data. Any data currently stored on a disk that appears in pvcsan output is destroyed when the HA admin node begins to run. As an alternative, you can move the data to another disk. Proceed with caution.

For example, the preceding commands indicate the following:

- /dev/sdc is recognized by both physical nodes.
- /dev/sdd is not in use currently.

In this example environment, /dev/sdc is a safe choice for the common disk.

4. Identify the world wide name (WWN) of the disk you want to use for the HA admin node.

To identify the disk, use a combination of ls and grep commands. The following example shows the command that returns the WWN of the disk you chose:

```
# ls -l /dev/disk/by-id/ | grep wwn
lrwxrwxrwx 1 root root  9 Nov 10 13:26 wwn-0x60080e5000233c340000039f4d90ab57 -> ../../sdc
lrwxrwxrwx 1 root root 10 Nov 10 13:26 wwn-0x60080e5000233c340000039f4d90ab57-part1 -> ../../sdc1
lrwxrwxrwx 1 root root 10 Nov 10 13:26 wwn-0x60080e5000233c340000039f4d90ab57-part2 -> ../../sdc2
```

The preceding command returned information about the disk itself and two partitions. Use the WWN of the disk itself, not the disk partitions. In this example, the WWN for the disk is as follows:

```
0x60080e5000233c340000039f4d90ab57
```

Observe the ID. A later procedure requires you to specify this WWN in the sac-ha-initial-setup.conf file.

CAUTION: This procedure uses data from an example environment. Do not assume that your environment can yield the same results. In your environment, correct disk analysis is not likely to produce the same effect. Do not assume that the analysis of your environment will also lead you to select /dev/sdc as your HA admin node shared disk.

5. Erase the existing data on the shared disk.

NOTE: This step is destructive. If necessary, move the data from the shared disk to another disk at your site.

As the root user, enter the following commands from one of the physical admin nodes:

```
# parted /dev/sdX mklabel gpt
# dd if=/dev/zero of=/dev/sdX bs=512 count=16384
```

For *X*, specify the identifier for the disk you want to erase.

6. Proceed to the following:

   Enabling an input-output memory management unit (IOMMU)

Enabling an input-output memory management unit (IOMMU)

Procedure

1. Log into each of the physical admin nodes as the root user.
2. On each physical admin node, open the following file in a text editor:

   `/etc/default/grub`
3. Search for the following string in the file:

   `GRUB_CMDLINE_LINUX_DEFAULT`
4. Add `intel_iommu=on` to the end of the `GRUB_CMDLINE_LINUX_DEFAULT` line.
5. On each physical admin node, save and close the edited file.
6. On each physical admin node, enter one of the following commands:

   On RHEL systems, enter the following:

   ```
   # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
   ```

   On SLES systems, enter the following:

   ```
   # grub2-mkconfig -o /boot/grub2/grub.cfg
   ```
7. Proceed to the following:

   Verifying the configuration

Verifying the configuration

Procedure

1. Log into each of the physical admin nodes as the root user.
2. Enter the following command on each physical admin node to verify the time zone:

   ```
   # date
   ```
3. Enter the following command on each physical admin node to verify the hostnames and the IP addresses:

   ```
   # cat /etc/hosts
   ```
4. Enter the following command on each physical admin node to verify the time:

   ```
   # chronyc sources -v
   ```

5. Enter the following command on each physical admin node to verify the network configuration:

```
# ip addr
```

   In the output, verify that all interfaces are in the eth*X* format.

6. Use the hostnamectl command to verify that the static host is set.

   For example, in the following output, the first line is Static hostname: *name*. Make sure that the hostname you specified for the physical admin node is the one that appears in the *name* field. The output shows the hostname set correctly on physical node hikari.

```
# hostnamectl
   Static hostname: hikari
         Icon name: computer-server
           Chassis: server
        Machine ID: 68c22b359c3b486a8576088cc3538beb
           Boot ID: 1274c1d3b2884cacb90d368c616b2ed5
  Operating System: Red Hat Enterprise Linux 8.0 (Ootpa)
       CPE OS Name: cpe:/o:redhat:enterprise_linux:8.0:GA
            Kernel: Linux 4.18.0-80.el8.x86_64
      Architecture: x86-64
```

7. Enter the following command on each physical admin node to reboot:

```
# reboot
```

8. Repeat Step **2** through Step **5** to make sure that your configuration persisted through the reboot.

9. Enter the following command on each physical admin node to make sure that IOMMU is enabled:

```
# dmesg | grep -E "DMAR: IOMMU"
```

   The output is as follows on a correctly configured system:

```
[    0.000000] DMAR: IOMMU enabled
```

10. Proceed to the following:

    Creating and installing the HA software repositories on the physical admin nodes

Creating and installing the HA software repositories on the physical admin nodes

The following procedure explains how to install the software repositories on each node.

Procedure

1. Use the ssh command to log into one of the physical admin nodes.

2. Copy the installation files (the operating system .iso files) to /var/opt/sgi on the node.

   Each operating system requires at least one operating system .iso file, and some operating systems require additional .iso files. Review the tables in the following topic to make sure that you have all the software you need:

   Operating system releases supported in the HPE Performance Cluster Manager 1.3.1 release

3. Log into the other admin node, and copy the installation files to the other admin node.

   Complete the following steps:

- Use the ssh command to log into the other admin node.
- When prompted, provide the root user login and password credentials.
- Use the rsync command to copy the files from this admin node to the other admin node.

  For example, assume that you used ssh to log into a node named admin2. To copy the files from the node named admin1 to the node named admin2, enter the following command:

  ```
  # rsync -avz admin1:/var/opt/sgi/*.iso /var/opt/sgi/
  ```

4. Proceed to the following:

   - If this node is the first HA admin node, go back to Step **1** and repeat these steps on the other node.
   - If this node is the second HA admin node, proceed to the following:

     Configuring the storage unit

Preparing to run the HA admin node configuration script

The configuration setup script configures the two physical nodes to communicate with each other and the storage unit. Edit this script and provide information within the script before you run the script.

The following procedure explains how to edit the setup script and provide the information that the script requires.

Procedure

1. Decide which node you want to designate as physical node 1 and physical node 2.
2. Log into each of the physical nodes as the root user.

   Each physical node sees itself as the primary physical node. Each physical node sees the other node as the secondary physical node.
3. Open the following file on each node, add the local NTP servers at your site, and save the file:

   /etc/chrony.conf
4. On each node, enter the following command to restart chronyd:

   ```
   # systemctl restart chronyd
   ```
5. On physical node 1, enter the following command:

   ```
   ip addr | grep eth
   ```

   The command displays NIC and MAC addresses.

   For example, you need the bolded information in the output that follows:

   ```
   # ip addr | grep eth
   2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq master state UP group default qlen 1000
       link/ether ec:eb:b8:89:03:40 brd ff:ff:ff:ff:ff:ff
   3: eth3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen 1000
       link/ether ec:eb:b8:89:03:41 brd ff:ff:ff:ff:ff:ff
       inet 192.168.0.1/24 brd 192.168.0.255 scope global noprefixroute eth3
   4: eth4: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc mq state DOWN group default qlen 1000
       link/ether ec:eb:b8:89:03:42 brd ff:ff:ff:ff:ff:ff
   5: eth1: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master state UP group default qlen
   1000
       link/ether 48:df:37:66:c2:30 brd ff:ff:ff:ff:ff:ff
   ```

```
6: eth2: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 1500 qdisc mq master state UP group default qlen
1000
    link/ether 48:df:37:66:c2:30 brd ff:ff:ff:ff:ff:ff
7: eth5: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc mq state DOWN group default qlen 1000
    link/ether ec:eb:b8:89:03:43 brd ff:ff:ff:ff:ff:ff
    link/ether 48:df:37:66:c2:30 brd ff:ff:ff:ff:ff:ff
    link/ether 48:df:37:66:c2:30 brd ff:ff:ff:ff:ff:ff
    link/ether ec:eb:b8:89:03:40 brd ff:ff:ff:ff:ff:ff
    link/ether fe:54:00:f5:41:d7 brd ff:ff:ff:ff:ff:ff
    link/ether fe:54:00:d3:ff:0c brd ff:ff:ff:ff:ff:ff
```

The MAC addresses are the highlighted text in the preceding output. You need this information to populate the fields that appear in the cluster configuration file after the following string:

```
# Network MACs for physical machine 1
```

6. On physical node 2, enter the following command:

   ```
   ip addr | grep eth
   ```

   Observe the output.

   You need this information to populate the fields that appear in the cluster configuration file after the following string:

   ```
   # Network MACs for physical machine 2
   ```

7. On physical node 1, use a text editor to open the following file:

   ```
   /etc/opt/sgi/sac-ha-initial-setup.conf
   ```

   The software copies `sac-ha-initial-setup.conf` to `sac-ha-initial-setup.conf.example`. If you edit the file but subsequently discard your edits, reinstate the original file and remove the `.example` suffix.

8. Set the path to the `.iso` file for the admin node.

   You need this information for the `admin_iso_path=` variable.

   Enter the following commands:

   ```
   # mkdir /root/sw
   # ssh phys_admin2
   # mkdir /root/sw
   # scp host_system:/path/cm-admin-install-1.3.1-os-x86_64.iso /root/sw/
   # rsync -avz /root/sw/ phys_admin:/root/sw/
   # exit
   ```

   The variables are as follows:

   - For *host_system*, specify the name of the computer that currently hosts the `.iso` file.
   - For *path*, specify the path to the `.iso` file on the hosting computer.
   - For *os*, specify the name of the operating system.

   For example, if you downloaded the `.iso` file to a Linux laptop, the `scp` command might look as follows:

   ```
   # scp user1@desktop:/home/user1/iso/\
   cm-admin-install-1.3.1-rhel81-x86_64.iso /root/sw/
   ```

9. On physical node 1, complete the lines in the `sac-ha-initial-setup.conf` file that the software requires to be edited for this cluster.

   This file pertains to the two physical nodes for this HA admin node.

   The cluster configuration file contains several lines that end in `=""`. Some of these lines contain default settings that you must assess for your site. For the other lines that end in `=""`, specify information for your HA cluster. The file contains comments that provide guidance regarding how to complete each line. **Table 1: Configuration file inputs** shows the lines that you must edit within the file.

NOTE: The `sac-ha-initial-setup.conf` file contains many fields. You do not have to populate all the fields with information from your cluster.

Table 1: Configuration file inputs contains information about the fields that you must edit. Do not edit the other fields.

Table 1: Configuration file inputs

Configuration file line	Information to provide
Physical node 1 - MAC addresses:	
`phys1_eth0=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys1_eth1=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys1_eth2=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys1_eth3=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
Physical node 2 - MAC addresses:	
`phys2_eth0=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys2_eth1=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys2_eth2=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.
`phys2_eth3=""`	To retrieve this MAC address, use the output from the `ip addr show` command (earlier in this procedure) or obtain it from your network administrator.

Table Continued

Configuration file line	Information to provide
Physical node 2 - hostname and IP address:	
`phys2_house_hostname=""`	Obtain this hostname from your network administrator. Specify the hostname. Do not specify the FQDN.
`phys2_house_ip=""`	Obtain this IP address from your network administrator. This is the IP address on your house network for access to the second physical admin node.
Physical node 1 - BMC hostname and BMC IP address:	
`phys1_bmc_hostname=""`	Obtain the BMC hostname for physical node 1 from your network administrator.
`phys1_bmc_ipaddr=""`	Obtain the BMC IP address for physical node 1 from your network administrator.
Physical node 2 - BMC hostname and BMC IP address:	
`phys2_bmc_hostname=""`	Obtain the BMC hostname for physical node 2 from your network administrator.
`phys2_bmc_ipaddr=""`	Obtain the BMC IP address for physical node 2 from your network administrator.
Additional information:	
`wwn=""`	Specify the information you retrieved in the following procedure: **Configuring the storage unit** The value you need consists of a number string. For example: `wwn="60080e5000233c340000039f4d90ab57"`
`volume_group_ID=""`	The number of the LUN on the storage device. Typically, this value is `0`. For information about how to derive this number, see your storage unit documentation.
`admin_iso_path=""`	The full path to the installation `.iso` file. Specify the path that you configured in Step **8**.

Table Continued

Configuration file line	Information to provide
`cpus=""`	The number of virtual CPUs assigned to the virtual machine admin node that manages the cluster. The maximum number is *max-cpu_threads* - 4. The following command retrieves CPU information for the `cpus=` field in the configuration file: `# `**`less /proc/cpuinfo`** In the output, look for the values for the following fields: • `processor` • `cpu cores` To arrive at the correct specification for the `cpus=` field for your cluster, use the information in these fields. The comments in the file also contain information about how to specify this field. For example, set `cpus="6"` or `cpus="8"`.

`memory=""`	The amount of memory allocated to the virtual machine admin node that manages the cluster.
	The following command retrieves information for the `memory=` field in the configuration file:
	`# `**`free -h`**
	`             total . . .`
	`Mem:           62G . . .`
	`Swap:         2.0G . . .`
	In the output, observe the values under the `total` column for the `Mem` field and the `Swap` field.
	Typically, you can allocate 4GB per virtual CPU that you specified on the `cpus=` line. The comments in the file also contain information about how to specify this field.
	For example, if you specified `cpus="6"`, specify `memory="24GB"`. If you specified `cpus="8"`, specify `memory="32GB"`.

Table Continued

Configuration file line	Information to provide

`rootsize=""`	Specify the amount of shared disk space on the storage unit that you want to allocate to the admin node virtual machine.
	The `fdisk` command retrieves information for the `rootsize=` field in the configuration file. This command has the following format:
	`fdisk -l` *`disk`*
	For *disk*, specify the disk identifier for the shared disk.
	For example:
	`#` **`fdisk -l /dev/sdb`**
	`Disk /dev/sdb: 4000.0 GB, . . .`
	`.`
	`.`
	`.`
	The `rootsize=` field requires a value in MB and recommends that you specify a value that is 80% of the LUN size. The minimum size is 94GB.
	For this example output, calculate 4000 X 1024 X 0.8, which yields 3276800. For the configuration file, specify `rootsize=3276800`.

`mail_to="root@localhost"`	Email address of the `root` user.

NOTE: At this time, edit only the fields shown in the preceding table. The comments in the `sac-ha-initial-setup.conf` file describe other fields that you can set in a troubleshooting situation.

10. Save and close the `sac-ha-initial-setup.conf` file on physical node 1.

11. Proceed to the following:

Running the highly available (HA) admin node configuration script

Running the highly available (HA) admin node configuration script

The configuration script configures the two physical nodes to run together. To obtain help information for the `sac-ha-initial-setup.conf` script, enter the following command:

```
# sac-ha-initial-setup --help
```

The configuration script configures the admin nodes to work together, but the script does not configure the shared storage.

When you run the configuration script, one or more steps might fail. If a step fails, you can stop the script, correct the problem, and restart the script.

The following procedure explains how to run the configuration script.

Procedure

1. Log into the first (or primary) physical admin node as the root user.
2. Enter the following command to navigate to the root directory:

   ```
   # cd ~
   ```

3. (Optional) Start the `ssh` agent in a way that suppresses nonfatal messages about the authentication agent.

   For example, in the bash shell, enter the following command:

   ```
   # exec ssh-agent bash
   ```

   The system might issue an erroneous message of the following type, which you can safely ignore:

   ```
   Could not open a connection to your authentication agent.
   ```

   To suppress the erroneous message, run the example `exec` command now and after each boot that this installation process requires. After installation, the system no longer issues this message.

4. Enter the following command to configure the HA system:

   ```
   # sac-ha-initial-setup
   ```

5. As the script prompts, enter the following command on each physical admin node to reboot the nodes:

   ```
   # reboot
   ```

   Wait for the boot to complete.

6. Log back into both of physical admin nodes as the root user.
7. On each of the physical admin nodes, use the `df` command to verify that the `/images` file system is mounted.

   For example:

   ```
   # df -h /images

   Filesystem             Size  Used Avail Use% Mounted on
                          280G  259M  280G   1% /images
   ```

8. Run the configuration script again on physical admin node 1.

   For example, enter the following on physical admin node 1:

   ```
   # sac-ha-initial-setup
   ```

9. (Optional) Enter the following command to monitor the configuration on physical admin node 2 and to make sure that the `virt` resource started:

   ```
   # crm_mon
   Last updated: Thu Nov 10 21:11:11 2019
   Last change: Thu Nov 10 21:08:35 2019 by root via cibadmin on mici-admin
   Stack: classic openais (with plugin)
   Current DC: mici-admin2 - partition with quorum
   Version: 1.3.12-f47ea56
   2 Nodes configured, 2 expected votes
   11 Resources configured
   Online: [ mici-admin mici-admin2 ]

   p_ipmi_fencing_1       (stonith:external/ipmi):        Started mici-admin2
   p_ipmi_fencing_2       (stonith:external/ipmi):        Started mici-admin
   sbd_stonith    (stonith:external/sbd): Started mici-admin
   Clone Set: dlm-o2cb-fs-images-clone [dlm-o2cb-fs-images-group]
      Started: [ mici-admin mici-admin2 ]
   ```

```
virt   (ocf::heartbeat:VirtualDomain): Started mici-admin2
mailTo (ocf::heartbeat:MailTo):        Started mici-admin2
```

10. Verify that the admin node image was created with the size you specified.

    For example:

    ```
    # ls -lh /images/vms/
            total 129G
            -rw------- 1 qemu qemu 128G May 13 16:08 sac.img
    ```

    NOTE: For information about HA admin node configuration troubleshooting, see the following:

    Troubleshooting an HA admin node configuration

11. Proceed to the following:

    Starting the HA virtual manager and installing the cluster manager on the virtual machine

Starting the HA virtual manager and installing the cluster manager on the virtual machine

The following procedure explains how to bring up the virtual machine manager. The HA installation and configuration process creates a virtual machine. One of the two physical admin nodes hosts the virtual machine at any given time.

Procedure

1. Log into physical node 1.

   None of the previous HA configuration steps required a graphics terminal. You could complete all the previous steps from a text-based terminal. Starting with this procedure, however, you are required to log in from a graphics terminal, as follows:

   - Log into the physical console and make sure that the cluster is booted at run level 5 (`init` 5).
   - Log in from a remote terminal through an `ssh` session with X11 forwarding. For example:

     ```
     # ssh -C -XY root@physical_node1_addr
     ```

     For *physical_node1_addr*, specify the IP address or hostname of physical node 1.
   - Log in through a VNC session. Again, make sure that the cluster is booted at run level 5 (`init` 5).

2. On physical node 1, enter the following command:

   ```
   # virt-manager &
   ```

3. In the **Virtual Machine Manager** window, click **File > Add Connection**.
4. On the **Add Connection** popup, complete the following steps:

   - Click the **Connect to remote host** box.
   - In the **Hostname** field, enter the hostname of physical node 2.
   - Click the **Autoconnect** box.
   - Click **Connect**.

5. In the **Virtual Machine Manager** window, double click the **sac running** icon.

The window that appears is the interface to the virtual machine that runs on the physical nodes. This window is the interface to the admin node that you can use for system administration tasks.

6. Use the arrow keys to select **Install: Wipe Out and Start Over: Prompted**, and press Enter.

7. At the **Enter number of slots to allow space for: (1-10):** prompt, enter the integer number of slots you want on this cluster, and press Enter.

8. At the **Enter which slot to install to: (1-10):** prompt, enter the integer number that corresponds to the slot you want to install, and press Enter.

9. At the **Destructively bypass sanity checks? (y/n):** prompt, enter **y**, and press Enter.

10. At the **Is this a physical admin node in an SAC-HA configuration? (normally no) (y/n):** prompt, enter **n**, and press Enter.

11. At the **Use predictable network names for the admin node? (normally yes) (y/n):** prompt, enter **y**, and press Enter.

    For information about predictable network names, see the following:

    Predictable network interface card (NIC) names

12. At the **Additional parameters (like console=, etc):** prompt, complete the following steps:

    - Enter **`console=ttyS0,115200n8`**.
    - Press Enter.

    When you complete this step, you enable the ability to log into the virtual admin node from one of the physical admin nodes. For more information about how to log into the virtual admin node, see the following:

    Connecting to the virtual admin node in a cluster with a highly available (HA) admin node

13. At the **OK to Proceed? (y/n):** prompt, enter **y**, and press Enter.

14. Wait for the software to install on the virtual machine.

15. Configure an operating system and network for the virtual machine.

    Use the graphical connection to the virtual machine, and complete one of the following procedures:

    - **Configuring RHEL 8.X or RHEL 7.X on the admin node**
    - **Configuring SLES 15 SPX and SLES 12 SPX on the admin node**

    After this step is complete, there is no need to log into the physical hosts. Also, there is no need to use the `virt-manager` tool. To connect to the HA admin node, use one of the following commands to log into the virtual machine:

    - `ssh -C -XY root@`*`admin_vm_addr`*

      or

    - `ssh root@`*`admin_vm_addr`*

    For *admin_vm_addr*, enter the admin node virtual machine IP address or hostname.

NOTE: To install software from physical media onto a cluster with an HA admin node, use the `virt-manager` command. The `virt-manager` command lets you take the DVD drive from a physical admin node and attach the DVD drive to the virtual admin node. During the installation, but after the `crepo` commands are complete, make sure to detach the DVD drive. If you do not detach the DVD drive from the virtual admin node, the DVD drive can become a problem during a failover.

16. Proceed to the following:

Specifying the cluster configuration

Installing the cluster software on the admin node

Installing the cluster software includes the following actions:

- Creating repositories for software installation files and updates.
- Installing the admin node cluster software.
- Configuring the cluster domain and examine other network settings. The cluster domain is likely to be different from the site network domain on the admin node itself.
- Configuring the NTP server.
- Installing the cluster software infrastructure. This step can take 30 minutes.
- Configuring the house network domain name server (DNS) resolvers.
- (Optional) Configuring an external DNS.

The following topics explain how to install the cluster software on the admin node:

- **Specifying the cluster configuration**
- **(Conditional) Using the menu-driven cluster configuration tool to specify the cluster configuration**
- **Completing the admin node software installation**
- **(Optional) Configuring external domain name service (DNS) servers**

Specifying the cluster configuration

The following procedure explains how to use one of the following to configure the cluster:

- The cluster definition file

  Or

- The cluster configuration tool

Procedure

1. Locate the cluster manager software distribution DVDs, or verify the path to the online software repository at your site.

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

2. From the VGA screen, or through an `ssh` connection, log into the admin node as the root user, as follows:

   - For a single-node admin node, Hewlett Packard Enterprise recommends that you run the cluster configuration tool as follows:

     - From the VGA screen

       or

     - From an `ssh` session to the admin node

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

- For an HA admin node, create an `ssh` connection to the host that is running the `virt` resource, and enter the `virt-viewer` command. For example:

  ```
  # ssh -C -XY root@phys_admin1
  # virt-viewer
  ```

  If the Virtual Machine Manager interface does not appear, log into the physical node, and enter `virt-viewer sac` on that host.

3. (Conditional) Open the ports that the cluster manager requires.

Complete this step if you configured a firewall on the admin node or anywhere else in the cluster.

The cluster manager requires the following ports:

- External ports required for SSH: TCP 22
- (Conditional) External port required for Kibana: TCP 5601

  TCP port 5601 is required if you want to use Kibana to access the centralized log files.
- External ports required for webpage and GUI: TCP 80, 443, 1099, and 49150.

  It is possible to start the cluster manager web server on a different port. For more information, see the following:

  Starting the cluster manager web server on a non-default port
- Internal ports required for monitoring: UDP 48555 - 49587

To avoid monitoring failures, do not permit other software to use the cluster manager ports.

For more information about port requirements, see the following:

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

4. Decide how you want to specify information to the `configure-cluster` command and proceed:

- If you have a cluster definition file, use that file as input to the `configure-cluster` command. Complete the following steps:

  ◦ Use the `cm repo add` command, in the following format, to create a repository for the installation package:

    ```
    cm repo add path_to_iso
    ```

    For *path_to_iso*, specify the full path to installation ISO.

    If you have hard media mounted in the admin node DVD drive, specify the path to that media. If operating system and cluster manager software reside in an ISO file on your network, specify the path to the files on your network.

    For example, enter the following commands to add a repository for a SLES ISO file that is required for SLES platforms and verify the repositories:

    ```
    # cm repo add /tmp/SLE-15-SP1-Packages-x86_64-GM-DVD1.iso
    # cm repo show
    ```

  ◦ Enter the following command to define the cluster according to the content in the cluster definition file:

    ```
    # configure-cluster --configfile path
    ```

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

  - Proceed to the following:

    Completing the admin node software installation

- If you do not have a cluster definition file, proceed to the following to use the `configure-cluster` command to start the menu-driven cluster configuration tool:

  (Conditional) Using the menu-driven cluster configuration tool to specify the cluster configuration

(Conditional) Using the menu-driven cluster configuration tool to specify the cluster configuration

The cluster configuration tool presents you with many default settings. Hewlett Packard Enterprise recommends that you keep the default settings if possible.

Procedure

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

   ```
   # configure-cluster
   ```

2. On the **House Network Interface Selection** screen, do the following:

   - Use the spacebar and arrow keys to select the network interface card (NIC) you want to use for the cluster house network.

     Make sure that the NIC you select has the IP address that you want people to use when they log into the cluster admin node from an outside public network.

   - Click **OK**.

3. On the **Management Network Interfaces Selection** screen, do the following:

   - Use the spacebar and arrow keys to select one or two NICs for the management network.
   - Click **OK**.

4. On the screen that asks **Do you want to use a separate, dedicated NIC to handle BMC traffic on the Management Network?**, click **Yes** or **No**.

   If you clicked **No**, proceed to the next step in this procedure. When you click **No**, the cluster manager uses the NICs you selected in the previous step for BMC traffic.

   If you clicked **Yes**, the installer presents you with the **Management BMC Network Interfaces Selection** screen. Select one of the NICs on that screen for the separate BMC network, and click **OK**.

5. On the screen that asks **Choose Admin bonding mode used for the management network**, do the following:

   - Click **active-backup** or **802.3ad (LACP)**.

     These modes are as follows:

Mode	Effect
active-backup	Only one link in a bonded interface is active at a time. This mode requires no matching configuration on the management switch. Default.
802.3ad (LACP)	All links in a bonded interface are active at the same time. This mode requires that the Ethernet switch connected has matching LACP configuration on all links in the bonded interface. Hewlett Packard Enterprise recommends using this bonding mode when more than one interface connects to a management network on the admin node.

NOTE: If you configured a highly available (HA) admin node, select the bonding mode that you configured on the two physical admin nodes.

- On the **Main Menu** screen, click **OK**.

On a configured cluster, you can see the interfaces you specified in the following file:

```
/etc/opt/sgi/configure-cluster-ethernets
```

6. On the **Cluster Configuration Tool: Initial Cluster Setup** screen, select **OK** on the screen.

The message on the screen is as follows:

```
All the steps in the following menu need to be
completed in order.  Some settings are harder
to change once the cluster has been deployed.
```

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

This procedure guides you through the tasks to perform for each of the menu selections on the **Initial Cluster Setup** screen.

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

- The operating system software, either RHEL or SLES
- The cluster manager
- (Optional) HPE Message Passing Interface (MPI)

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

8. On the **One or more ISOs were embedded on the ...** screen, select **Yes**.

9. Wait for the software repositories to configure.

10. At the `press ENTER to continue` prompt, press **Enter**.

11. On the **You will now be prompted to add additional media ...** screen, select **OK**.

12. On the **Would you like to create repos from media? ...** screen, select one of the following:

- **Yes**. After you select **Yes**, proceed to the following:

Step **13**

- **No**. After you select **No**, proceed to the following:

  Step **15**

13. On the **Please either insert the media in your DVD drive ...** screen, select either **Inserted DVD** or **Use Custom path/url**.

Proceed as follows:

- To install the software from DVDs, perform the following steps:
  - Insert a DVD.
  - Select **Mount inserted DVD**.
  - On the **Media registered successfully with crepo ...** screen, select **OK**, and eject the DVD.
  - On the **Would you like to register media with SMC? ...** screen, select **Yes** if you have more software to register.

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

    If you select **No**, proceed to the next step.
- To install the software from a network location, perform the following steps:
  - Select **Use custom path/URL**.
  - On the **Please enter the full path to the mount point or the ISO file ...** screen, enter the full path in *server_name*:*path_name*/*iso_file* format. This field also accepts a URL or an NFS path. Select **OK** after entering the path.
  - On the **Media registered successfully with crepo ...** screen, select **OK**.
  - On the **Would you like to register media with SMC? ...** screen, select **Yes** if you have more software that you to register.

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

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

14. Repeat the following steps until all software is installed:

- Step **12**
- Step **13**

If you plan to configure MPT and run MPT programs, make sure to install the HPE Message Passing Interface (MPI) software.

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

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

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

17. On the **About to create secrets ...** popup window, select **Yes**.

18. On the **Admin node network and database will now be initialized** popup, select **OK**.

19. (Conditional) Create one or more data networks.

Complete this step if you have additional NICs that you want to use for data networks on non-ICE compute nodes.

To configure a data network, complete the following steps:

- On the **Cluster Network Settings** screen, select **A Add Subnet**, and select **OK**.
- On the **Select network type** screen, press the spacebar to move the asterisk (`*`) to the second line. This action selects the lower line, which now appears as follows:

  ```
  (*) 2 data
  ```

- Select **OK**.
- On the **Insert network name, subnet and netmask** screen, type in the information to define the data network. Use the arrow keys to move from field to field on this screen. Enter the following information:

Field	Information needed
name	A unique name for this network. For example, `data10g`.
subnet	The network IP address (start of the range) for the nodes on the data network.
netmask	The network mask for the nodes on routed management network.

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

You can specify the data network information in one of the following ways:

- In the cluster configuration file

  or

- As parameters to the `discover` command

20. On the **Cluster Network Settings** screen, select **S List and Adjust Subnet Addresses**, and select **OK**.

21. Verify the information on the **Caution: You can adjust ...** screen, and click **OK**.

22. Review the settings on the **Subnet Network Addresses - Select Network to Change** screen, and modify these settings only if necessary.

This screen displays the default networks and netmasks that reside within the cluster. Complete one of the following actions:

- To accept the defaults, select **Back**.
- To change the network settings, complete the following steps:
  - Highlight the setting you want to change, and select **OK**.
  - Enter a new IP address, and select **OK**.
  - Press Enter.

For example, it is possible that your site has existing networks or conflicting network requirements. For additional information about the IP address ranges, see the following:

Subnetwork information

On the **Update Subnet Addresses** screen, the **Head Network** field shows the admin node IP address. Hewlett Packard Enterprise recommends that you do not change the IP address of the admin node or leader nodes if at all possible. You can change the IP addresses of the InfiniBand network or the Omni-Path network. These networks are named **IB0** and **IB1**. You can change the `IB0` and `IB1` IP addresses to match the IP requirements of the house network, and then select **Back**.

23. On the **Cluster Network Settings** screen, select **D Configure Cluster Domain Name**, and select **OK**.
24. On the **Please enter the domain name for this cluster** pop-up window, enter the domain name, and select **OK**.

    The domain you specify becomes a subdomain of your house network.

    For example, enter `ice.americas.hpe.com`.
25. On the **Domain name configured** screen, click **OK**.
26. On the **Please adjust the domain_search_path as needed ...** screen, click **OK**.
27. Select **P Domain Search Path** to verify the domain search path.
28. In the **Please enter the Domain Search Path for this cluster** box, verify the information, adjust if needed, and click **OK**.
29. On the **Domain Search Path Configured** screen, click **OK**.
30. (Optional) On the **Cluster Network Settings** screen, select **U Configure Udpcast Settings**, and select **OK**.

    On the **Udpcast Settings** screen, select one of the following, and select **OK**.

    The selections are as follows:

    - **U Admin Udpcast RDV Multicast Address**
    - **T Admin Udpcast TTL**
    - **G Global Udpcast RDV Multicast Address**

    For each of the preceding selections, enter a value, and click **OK**.

    For information about the actions available from the preceding settings, select the setting. An informational window appears. When finished, click **Back** until you get to the **Cluster Network Settings** screen.
31. (Optional) On the **Cluster Network Settings** screen, adjust the VLAN settings.

    Depending on the cluster hardware and your site requirements, you might want to adjust the VLAN settings.

    - Adjusting the VLAN settings for an HPE Apollo 9000 cluster:

      If the cluster is an HPE Apollo 9000 with chassis management controllers (CMCs), you can disable or adjust the auto-generated routed VLANs on the management network. Complete the following steps:

      a. On the **Cluster Network Settings** screen, select **M Configure Management Network VLAN Settings**, and select **OK**.

      b. On the **Setting the CMCs per Management ...** screen, click **OK**.

      c. On the **Management VLAN Settings** screen, select one of the following settings, specify a value, and click **OK**.

        - **Management VLAN Start: # (default: 2001)**
        - **Management VLAN End: # (default: 2999)**
        - **CMCs per Management VLAN: # (default 8, 0=feature disabled)**

    d. When finished specifying new values, click **OK**.

    e. When all values are set, click **Back**.

- Adjusting the settings for an HPE SGI 8600 cluster:

  If the cluster has HPE SGI 8600 hardware with CMCs, rack leader controllers (RLCs), and cooling hardware, you can adjust VLAN numbers used by the cluster. Complete the following steps:

    a. On the **Cluster Network Settings** screen, select **V Configure HPE SGI 8600 VLAN Settings**, and select **OK**.

    b. On the **Warning: Changing VLAN settings ...** screen, click **OK**.

    c. On the **HPE SGI 8600 VLAN Settings** screen, select one of the following settings, specify a value, and click **OK**.

        - **Rack VLAN Start: # (default: 101)**
        - **Rack VLAN End: # (default: 1100)**
        - **Mcell VLAN: # (default 3)**
        - When finished, click **Back**.

    d. When finished specifying new values, click **OK**.

    e. When finished, click **Back**.

- Adjusting the settings for a cluster with multiple VLANs that requires L3 routing.

  If the cluster is configured to use multiple VLANs and it requires L3 routing to achieve end-to-end connectivity, you can adjust the settings. Use the following steps to change the VLAN numbers used by the supported routing protocols. The supported protocols are OSPF and routing information protocol (RIP).

    a. On the **Cluster Network Settings** screen, select **X Configure OSPF VLAN Settings**, and select **OK**.

    b. On the **OSPF VLAN #** screen, use the up and down arrows to highlight the field you want to specify, specify a value, and click **OK**:

        - **OSPF VLAN #: # (default:1999)**
        - **OSPF Base Network: <valid IP network address> (default 1.1.0.0)**
        - **OSPF Base Netmask: <valid IP subnet mask> (default 255.255.0.0)**
        - When finished specifying new values, click **OK**.
        - When finished, click **Back**.

    c. On the **Cluster Network Settings** screen, select **Y Configure RIP VLAN Settings**, and select **OK**.

    d. On the **Change RIP VLAN #** screen, use the up and down arrows to highlight the field you want to specify, specify a value, and click **OK**:

    - **RIP VLAN #: # (default:1998)**
    - **RIP Base Network: <valid IP network address> (default 1.2.0.0)**
    - **RIP Base Netmask: <valid IP subnet mask> (default 255.255.0.0)**
    - When finished specifying new values, click **OK**.
    - When finished, click **Back**.

  e. When finished, click **Back**.

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

33. On the **Initial Cluster Setup Tasks -- all Required** popup, select **S Perform Initial Admin Node Infrastructure Setup**, and select **OK**.

34. On the following screen, select **OK**:

```
A script will now perform the initial cluster
set up including setting up the database and
some network settings.
```

35. On the **Enter up to three DNS resolvers IPs** screen, make adjustments if needed, and select **OK**.

36. On the **Setting DNS Forwarders to X.X.X.X** screen, review the display and take one of the following actions:

- To change the display, select **No**, and make adjustments if needed.
- If the display is correct, select **Yes**.

37. On the **Copy admin ssh configuration ...** screen, take one of the following actions:

- To change the display, select **No**, and make adjustments if needed.
- If the display is correct, select **Yes**.

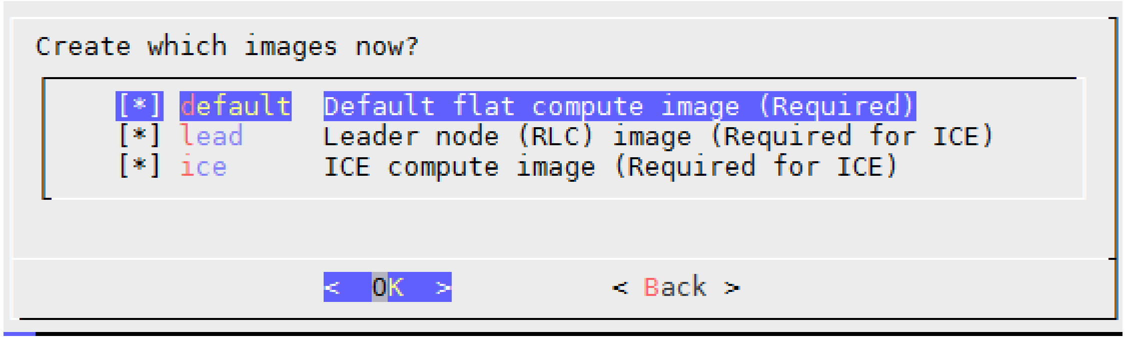
38. On the **Create which images now?** screen, confirm the images that you want to create.

The various cluster types need the following images:

Images needed on a cluster without leader nodes	Images needed on a cluster with ICE leader nodes	Images needed on a cluster with scalable unit (SU) leader nodes
`default`	`default`	`default`
	`lead`	
	`ice`	

NOTE: If you plan to configure scalable unit (SU) leader nodes, do not configure SU leader node images at this time. A later procedure explains how to create SU leader node images.

Figure 2: Create images screen shows an example. If the cluster does not have ICE leader nodes, use arrow keys and the spacebar to clear the fields **lead** and **ice**.



```
Create which images now?

    [*] default   Default flat compute image (Required)
    [*] lead      Leader node (RLC) image (Required for ICE)
    [*] ice       ICE compute image (Required for ICE)
```

Figure 2: Create images screen

Select **OK** when the screen shows the images that you want to create. It can take up to 30 minutes to create the images.

If you clear any fields, the installer does not create an image for that particular node type.

Wait for the completion message. The script writes log output to the following log file:

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

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

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

40. On the **Initial Cluster Setup Complete** screen, select **OK**.

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

41. On the **Initial Cluster Setup Tasks -- All Required** screen, select **M Configure Switch Management Network**, and click **OK**.

42. On the **Default Switch Management Network setting for newly discovered ...** screen, select **Yes** and select **OK**.

43. On the **Initial Cluster Setup Tasks -- All Required** screen, select **O Configure Monitoring**, and click **OK**.

The installation process installs and configures native HPE Performance Cluster manager monitoring software and Ganglia software on the cluster nodes. This step explains how to enable the monitoring software at installation time. You can enable various types of monitoring. By default, monitoring software is installed but not enabled.

To enable native monitoring, complete the following steps:

a. On the **Cluster Monitoring Settings** screen, select **Native Monitoring**, and click **OK**.

b. On the **Enable native monitoring?** screen, select **Y yes**, and click **OK**.

c. On the **Native monitoring has been set to enable** screen, click **OK**, and wait while the system configures native monitoring.

d. On the **Cluster Monitoring Settings** screen, click **Back**.

To enable Ganglia monitoring, complete the following steps:

a. On the **Cluster Monitoring Settings** screen, select **Ganglia Monitoring**, and click **OK**.

b. On the **Enable Ganglia Monitoring?** screen, select **Y yes**, click **OK**, and wait while the system configures Ganglia.

c. On the **Ganglia monitoring has been set to enable** screen, click **OK**.

d. On the **Cluster Monitoring Settings** screen, click **Back**.

To enable Nagios monitoring, complete the following steps:

a. On the **Cluster Monitoring Settings** screen, select **Nagios Monitoring**, and click **OK**.

b. On the **Enable Nagios Monitoring?** screen, select **Y yes**, click **OK**, and wait while the system configures Nagios.

c. On the **Nagios monitoring has been set to enable** screen, click **OK**.

d. On the **Cluster Monitoring Settings** screen, click **Back**.

To enable Kafka, Elasticsearch, and Alerta monitoring, complete the following steps:

a. On the **Cluster Monitoring Settings** screen, select **Kafka/ELK/Alerta Monitoring**, and click **OK**.

b. On the **Enable Kafka/ELK/Alerta Monitoring?** screen, select **Y yes**, click **OK**, and wait while the system configures Kafka, ELK, and Alerta services.

c. On the **Kafka/ELK/Alerta monitoring has been set to enable** screen, click **OK**.

d. On the **Cluster Monitoring Settings** screen, click **Back**.

To enable, start, stop, or disable monitoring after the cluster is running, use the `cm monitoring` command.

44. On the **Initial Cluster Setup Tasks -- All Required** screen, select **P Predictable Network Names**, and select **OK**.

45. On the **Default Predictable Network Names ...** popup, select **Yes** or **No**. These selections have the following effect:

    - Select **Yes** and select **OK** to use predictable names on future equipment. For example, if you select **Yes** here, the cluster is configured to add new equipment with predictable names later.

      If the admin node is configured with predictable names, this popup has **Yes** highlighted because that is the clusterwide default.

      Or

    - Select **No** and select **OK** to use legacy names on future equipment.

      If the admin node is configured with legacy names, this popup has **No** highlighted because that is the clusterwide default.

NOTE: Hewlett Packard Enterprise recommends that you do not mix predictable names with legacy names in the same cluster. To change the naming scheme for a cluster component, run the `discover` command (again) on that component, which reconfigures the component into the cluster with the alternative naming scheme. For more information about predictable names and legacy names, see the following:

Predictable network interface card (NIC) names

46. Select **Back**.

47. Select **Quit**.

48. Proceed to the following:

    Completing the admin node software installation

Completing the admin node software installation

The following procedure completes the admin node software installation.

Procedure

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

   For example, the following output is generated on a cluster with ICE leader nodes and with liquid cooling cells. If your system does not include liquid cooling cells, the `mcell_network` value displays `no`. Example output is as follows:

   ```
   # cattr list -g
   global
     blade_image_default     : ice-sles15sp1 4.4.73-5-default nfs
     ospf_vlan_tag           : 1999
     switch_mgmt_network     : yes
     blademond_scan_interval : 120
     udpcast_mcast_rdv_addr  : 224.0.0.1
     max_rack_irus           : 16
     rack_vlan_start         : 101
     predictable_net_names   : yes
     udpcast_receive_timeout : 5
     admin_udpcast_portbase  : 9000
     redundant_mgmt_network  : yes
    dhcp_bootfile           : grub2
     rack_vlan_end           : 1100
     conserver_ondemand      : no
     udpcast_retries_until_drop : 15
     discover_skip_switchconfig : no
     udpcast_max_bitrate     : 900m
     ospf_vlan_netmask       : 255.255.0.0
     udpcast_min_wait        : 10
     rip_vlan_netmask        : 255.255.0.0
     house_dns_servers       : 100.162.237.211; 100.162.236.210; 100.149.32.11;
     head_vlan               : 1
     rip_vlan_tag            : 1998
     udpcast_rexmit_hello_interval : 0
     mcell_network           : yes
     edns_udp_size           : 512
     ospf_vlan_network       : 1.1.0.0
     rip_vlan_network        : 1.2.0.0
     domain_search_path      : ib0.example.americas.sgi.com,example.americas.sgi.com,americas.sgi.com,engr.sgi.com,sgi.com
     cluster_domain          : example.americas.sgi.com
     udpcast_max_wait        : 10
     image_push_transport    : udpcast
     udpcast_portbase        : 43124
     mcell_vlan              : 3
     mgmt_vlan_end           : 2500
     mgmt_vlan_start         : 2001
     udpcast_min_receivers   : 1
     conserver_logging       : yes
     udpcast_ttl             : 1
   ```

   NOTE: The preceding output differs from cluster to cluster depending on configuration choices and hardware.

   To respecify any global values, start the cluster configuration tool again, and correct your specifications. To start the cluster configuration tool, enter the following command:

   ```
   # configure-cluster
   ```

2. (Conditional) Allocate the IP addresses used by the physical admin nodes for the private network within the cluster.

   Complete this step if you are configuring an HA admin node.

   a. Obtain the following values from the `sac-ha-initial-setup.conf` file:

      - `phys1_head_ip=`
      - `phys1_eth1=`

- `phys2_head_ip=`
- `phys2_eth1=`

b. Use the `discover` command in the following format to add the first physical admin node to the cluster:

```
discover --node 500,generic,mgmt_net_name=head,hostname1=physadmin1,\
mgmt_net_ip=phys1_head_ip_value,mgmt_net_macs=phys1_eth1_value
```

The variables are as follows:

- For *phys1_head_ip_value*, specify the value in the `sac-ha-initial-setup.conf` file for `phys1_head_ip`.
- For *phys1_eth1_value*, specify the value in the `sac-ha-initial-setup.conf` file for `phys1_eth1`

NOTE: The values of `500` and `physadmin1` in the preceding command can be any values. The `500` is the node number; for this value, pick a large value that is greater than the number of physical non-ICE compute nodes you ever expect to have in the cluster.

c. Use the `discover` command in the following format to add the second physical admin node to the cluster:

```
discover --node 501,generic,mgmt_net_name=head,hostname1=physadmin2,\
mgmt_net_ip=phys2_head_ip_value,mgmt_net_macs=phys2_eth1_value
```

The variables are as follows:

- For *phys2_head_ip_value*, specify the value in the `sac-ha-initial-setup.conf` file for `phys2_head_ip`.
- For *phys2_eth1_value*, specify the value in the `sac-ha-initial-setup.conf` file for `phys2_eth1`

NOTE: The values of `501` and `physadmin2` in the preceding command can be any values. The `501` is the node number; for this value, pick a large value that is greater than the number of physical non-ICE compute nodes you ever expect to have in the cluster.

d. Enter the following command to verify these values in the `/etc/hosts` file:

```
# cat /etc/hosts | grep physadmin
172.23.200.1  physadmin1.head.cm.cluster.net physadmin1 service500
172.23.200.2  physadmin2.head.cm.cluster.net physadmin2 service501
```

3. (Optional) Configure an unsupported switch.

If the cluster contains any unsupported switches, see the following topic and return here when finished:

Configuring a cluster that uses an unsupported Ethernet switch

4. Proceed to one of the following:

- To configure one or more external Domain Name Service (DNS) servers, which is an optional step, proceed to the following:

(Optional) Configuring external domain name service (DNS) servers

- To configure an HA ICE leader node, proceed to the following:

  Configuring a high availability (HA) ICE leader node

- If you do not want to configure an HA ICE leader node and do not want to configure an external DNS, proceed to the following:

  Updating the software repository

(Optional) Configuring external domain name service (DNS) servers

Perform the procedure in this topic to enable network address translation (NAT) gateways for the cluster. When external DNS and NAT are enabled, the host names for the nodes in the cluster resolve through external DNS servers. The nodes must be able to reach your house network.

NOTE: If you want to enable NAT, complete the procedure in this topic at this time. You cannot complete the procedure in this topic after you run the `discover` command. If you attempt to configure this feature after you run the `discover` command, the IP addresses assigned previously on the configured nodes remain.

Procedure

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

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

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

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

   ```
   # configure-cluster
   ```

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

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

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

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

8. Proceed to one of the following:

   - To configure an HA ICE leader node, proceed to the following:

     Configuring a high availability (HA) ICE leader node

   - If you do not want to configure an HA ICE leader node, proceed to the following:

     Updating the software repository

   NOTE: After the cluster is configured, use the procedure in the following topic to complete the NAT configuration:

   Configuring ICE compute nodes to use a non-ICE compute node as a network address (NAT) gateway (HPE SGI 8600 clusters)

Configuring a high availability (HA) ICE leader node

Complete the procedures in this chapter if you want to configure an HA ICE leader node.

When you configure an HA ICE leader node, the active HA ICE leader node is configured with the following:

- A second IP address on the following networks: head, gbe, and bmc. The other nodes in the cluster use these IP addresses to communicate to the HA ICE leader node. Both HA ICE leader nodes have an IP in the head-bmc network to facilitate communication with the baseboard management controller (BMC) of the other node.
- The main services to control the rack. For example, these services include the following: the hardware event tracker (HET), icerackd, clmgr-power, blademond, conserver, dhcpd, flamethrowerd, smc-leaderd, si_netbootmond, and the NFS server.

The two HA ICE leaders in the HA configuration have a direct Ethernet connection to each other. Corosync uses both of the following:

- bond0, which holds the interfaces of the management network.

  If bond0 goes down, there is still a connection between the two HA ICE leader nodes through the dedicated HA network. This situation could arise, for example, due to problems with the management switch to which the HA ICE leader nodes connect.
- The dedicated HA connection between the nodes.

Fencing is done using the STONITH (shoot-the-other-node-in-the-head) approach.

An IPMI plugin controls power-on actions and power-off actions.

The partition reserved for the current slot of each HA ICE leader node is composed of the following two volumes:

- A volume for the root (/) partition.
- A volume for the shared partition, which uses DRBD. The DRBD volume provides the file system for the blades.

  On the active node, the shared (DRBD) partition is mounted at /var/lib/sgi. The installer writes the ICE compute miniroot, per-host, and root file system, directories to this mount point. During a failover, the failing active node unmounts the DRBD partition and the other node mounts the DRBD partition as it takes over. The same thing happens when a user migrates the active HA ICE leader node function from one node to the other.

  You can use the cadmin command to configure the percentage of the slot to use for the root volume versus the shared volume.

The procedures in this chapter explain the following:

- How to evaluate the shared root volume default configuration for use at your site
- How to prepare the HA ICE leader node repositories
- How to verify the cluster definition file.

This chapter includes the following topics:

- **Specifying the root volume percentages**
- **Updating the cluster definition file**

NOTE: Complete the procedures in this chapter only if you are configuring two ICE leader nodes to work together as an HA ICE leader node.

Specifying the root volume percentages

An HA ICE leader node slot includes a volume for the root partition and a volume for the shared partition. You can specify the following:

- The amount of space in the slot that you want to allocate to the root partition.
- The amount of space in the slot that you want to allocate to the shared partition.

When the `discover` command runs, it creates volumes with the percentages you specified. The following are the default settings:

- 50% of the slot is allocated for the root volume.
- Of the space allocated for the root volume, 100% of that space is allocated for the shared volume.

For example, assume that you have a 100 GB slot. By default, 50 GB is allocated for the root volume when the `discover` command runs. Also by default, 100% of that remaining 50 GB is allocated for the shared volume when the `discover` command runs.

You can use the following cluster software commands to retrieve information about the current settings and to set site-specific percentages:

- `cadmin --show-ha-rlc-root-volume-size`
- `cadmin --set-ha-rlc-root-volume-size` *`percent`*. For *percent*, specify a whole integer. The default is 90. 10 is the lowest integer you can specify, and 90 is the highest integer you can specify.
- `cadmin --show-ha-rlc-shared-volume-size`
- `cadmin --set-ha-rlc-shared-volume-size` *`percent`*. For *percent*, specify a whole integer. The default is 100. 10 is the lowest integer you can specify, and 100 is the highest integer you can specify. If you have many compute node images, Hewlett Packard Enterprise recommends that you set this value to a high number.

Procedure

1. Use `cadmin` commands to query the allocation percentages.

   For example:

   ```
   # cadmin --show-ha-rlc-root-volume-size
   50%FREE
   # cadmin --show-ha-rlc-shared-volume-size
   100%FREE
   ```

2. (Optional) Reset the allocation percentages.

   Complete this step to reset the allocation percentages.

   For example, the following commands reset both percentages:

   ```
   # cadmin --set-ha-rlc-root-volume-size 90
   # cadmin --set-ha-rlc-shared-volume-size 80
   ```

3. Proceed to one of the following:

- **Preparing the images for RHEL 8.X or RHEL 7.X HA ICE leader nodes**
- **Preparing the images for SLES 15 or SLES 12 HA ICE leader nodes**

Preparing the images for RHEL 8.X or RHEL 7.X HA ICE leader nodes

As part of the HA ICE leader node enablement procedure, this topic explains how to add the following to the cluster:

- The RHEL HA software and updates, if any
- The RHEL Resilient Software and updates, if any

The instructions in this topic assume that you get your updates directly from the operating system vendor, from a web location at your site, or from a server at your site. If your site uses hard media, you can adapt the instructions in this topic to include mounting media.

The following procedure explains how to create the system software repository for RHEL ICE leader nodes.

NOTE: Some of the output examples in the following procedure have been wrapped for inclusion in this documentation.

Procedure

1. (Conditional) Download the RHEL ISO from Red Hat, Inc., onto the admin node.

   Complete this step if the ISO was added from a remote repository. For example, from a web-based repository. If you downloaded an ISO to the admin node, you do not need to perform this step.

2. (Conditional) Add RHEL HA and RHEL Resilient Storage repository updates.

   Complete this step if the ISO was added from a remote repository. If you downloaded an ISO to the admin node, you do not need to perform this step.

   Use the following command format:

   ```
   cm repo add --custom repo_name web_location
   ```

   For *repo_name*, specify the name of the repository to update. In this case, update the following:

   - `RHEL81-HA`
   - `RHEL81-ResilientStorage`
   - `RHEL8-HA`
   - `RHEL8-ResilientStorage`

   For *web_location*, specify the URL path to the repository. If a server at your site hosts operating system updates, specify that path here.

   For example, the following commands add the latest updates to RHEL 8.1:

   ```
   # cm repo add --custom RHEL8-baseos \
   http://server-with-remote-repos/rhel-8-for-x86_64-baseos-rpms/
   # cm repo add --custom RHEL8-appstream \
   http://server-with-remote-repos/rhel-8-for-x86_64-appstream-rpms/
   ```

   In this example, the commands access the software from a remote repository. At your site, replace the `server-with-remote-repos` with a valid path to a server at your site.

3. Select the repository updates:

   ```
   cm repo select repo_name
   ```

   For *repo_name*, specify the name of the repository.

   For example, the following command selects the RHEL 8.1 updates:

   ```
   # cm repo select RHEL8-baseos
   # cm repo select RHEL8-appstream
   ```

4. (Optional) Verify the updates.

   For example:

   ```
   # cm repo show
   * RHEL8-baseos
   * RHEL8-appstream
   ```

5. Display the current repositories and verify that the new repositories exist.

   For example:

   ```
   # cm repo show
   * RHEL8-baseos http://server-with-remote-repos/rhel-8-for-x86_64-baseos-rpms/
   * RHEL8-appstream http://server-with-remote-repos/rhel-8-for-x86_64-appstream-rpms/
   * Cluster-Manager-1.3.1-rhel8.1 : /opt/clmgr/repos/cm/Cluster-Manager-1.3.1-rhel8.1
   * Red-Hat-Enterprise-Linux-8.1 : http://server-with-updates/rhel8.1
   ```

6. Clone the ICE leader node image, and verify that the new clone exists.

   For example, the following command clones image `lead-rhel8.1` to image `lead-rhel8.1-ha`:

   ```
   # cm image copy -s lead-rhel8.1 -i lead-rhel8.1-ha
   ```

   The following output shows the new clone:

   ```
   # cm image show
   ...
   lead-rhel8.1-ha
   ...
   ```

7. Prepare the admin node for the HA ICE leader node configuration.

   ```
   # cm node update -n admin
   # cm node dnf -n admin install ha-rlc-admin
   ```

8. Prepare the ICE leader node image for the HA ICE leader node configuration.

   For example, if the ICE leader node image is called `lead-rhel8.1-ha`, enter the following commands:

   ```
   # cm image update -i lead-rhel8.1-ha
   # cm node dnf -i lead-rhel8.1-ha install ha-rlc-lead
   ```

9. Proceed to the following:

   Updating the cluster definition file

Preparing the images for SLES 15 or SLES 12 HA ICE leader nodes

As part of the HA ICE leader node enablement procedure, this topic explains how to add the following to the cluster:

- The SLES 15 SPX or SLES 12 SPX HA software
- The SLES 15 SPX or SLES 12 SPX HA software updates, if any
- The SLES 15 SPX or SLES 12 SPX software updates, if any

The instructions in this topic assume that you get your updates directly from the operating system vendor, from a web location at your site, or from a server at your site. If your site uses hard media, you can adapt the instructions in this topic to include mounting media.

The following procedure explains how to create the system software repository for HA ICE leader nodes on SLES 15 SPX. Use your SLES documentation if the cluster runs SLES 12 SPX.

Procedure

1. Download the SLES ISO from SUSE onto the admin node.

2. Add the SLES 12 software to the cluster.

   Complete this step if the cluster runs SLES 12. The SLES 15 operating system does not require you to complete this step.

   For SLES 12 SP5, enter the following commands:

   ```
   # cm repo add /path/SLE-12-SP5-HA-DVD-x86_64-GM-CD1.iso
   # cm repo select SUSE-Linux-Enterprise-High-Availability-Extension-12-SP5
   ```

   For *path*, specify the path to the ISO image at your site.

3. Add new repos to the cluster.

   ```
   # cm repo add --custom "sles15sp1-ha-updates" \
   http://updateserver.example.com/SLE-Product-HA/15-SP1/x86_64/update/
   # cm repo add --custom "sles15sp1-baseos" \
   http://updateserver.example.com/SLE-Module-Basesystem/15-SP1/x86_64/update/
   # cm repo add --custom "sles15sp1-python2" \
   http://updateserver.example.com/SLE-Module-Python2/15-SP1/x86_64/update/
   # cm repo add --custom "sles15sp1-moduledevelopment" \
   http://updateserver.example.com/SLE-Module-Development-Tools/15-SP1/x86_64/
   update/
   ```

4. Select the new repositories.

   ```
   # cm repo select sles15sp1-ha-updates
   # cm repo select sles15sp1-baseos
   # cm repo select sles15sp1-python2
   # cm repo select sles15sp1-moduledevelopment
   ```

5. Display the current repositories, and verify that the new repositories exist.

   For example:

   ```
   # cm repo show
   sles15sp1-ha-updates : /opt/clmgr/repos/cm/sles15sp1-ha-updates
   sles15sp1-baseos : /opt/clmgr/repos/cm/sles15sp1-baseos
   sles15sp1-python2 : /opt/clmgr/repos/cm/sles15sp1-python2
   sles15sp1-moduledevelopment : /opt/clmgr/repos/cm/sles15sp1-moduledevelopment
   Cluster-Manager-1.3.1-sles15sp1 : /opt/clmgr/repos/cm/Cluster-Manager-1.3.1-
   sles15sp1
   SUSE-Linux-Enterprise-Server-15-SP1 : /opt/clmgr/repos/distro/sles15sp1
   HPE-MPI-1.6-sles15sp1 : /opt/clmgr/repos/cm/HPE-MPI-1.6-sles15sp1
   ```

6. Clone the ICE leader node image, and verify that the new clone exists.

For example, the following commands clone image `lead-sles15` to image `lead-sles15-ha` and then show the new clone:

```
# cm image copy -s lead-sles15sp1 -i lead-sles15sp1-ha
# cm image show
...
lead-sles15sp1-ha
...
```

7. Read the SLES license agreement.

   After you read the license agreement, you can use the `--auto-agree-with-licenses` parameter on the `cm image zypper` command.

8. Prepare the admin node for the HA ICE leader node configuration:

   ```
   # cm node update -n admin
   # cm image zypper -n admin install --auto-agree-with-licenses ha-rlc-admin
   ```

9. Prepare the ICE leader node image for the HA ICE leader node configuration.

   For example, if the ICE leader node image is called `lead-sles15sp1`, enter the following commands:

   ```
   # cm image update -i lead-sles15sp1-ha
   # cm node zypper -i lead-sles15sp1-ha install \
   --auto-agree-with-licenses ha-rlc-lead
   ```

10. Proceed to the following:

    Updating the cluster definition file

Updating the cluster definition file

The topic in this procedure explains how to ensure that the cluster definition file includes the correct information for the HA ICE leader nodes.

The following procedure explains how to verify the cluster definition file.

Procedure

1. Open the cluster definition file.

   This file can reside anywhere. For example, during manufacturing, HPE writes the cluster definition file to the following location:

   `/var/tmp/mfgconfigfile`

2. Make sure that the `ha=` field is set correctly.

   For example, to configure all the ICE leader nodes as HA ICE leader nodes, make sure the `ha` field is set.

   For example, you might see the following settings in the cluster definition file:

   - `ha=0`, which specifies that the ICE leader nodes are individual nodes and no high availability is configured.
   - `ha=all`, which specifies that you want to install the cluster manager so that all ICE leader nodes are HA ICE leader nodes. Hewlett Packard Enterprise recommends that you specify `ha=all` if possible.
   - `ha=1`, which specifies that you want to configure HA ICE leader node 1, which consists of two individual nodes. Use the `ha=` field this way in the following situation:

    - Your plan is to install (or reinstall) the cluster manager software on HA ICE leader node 1 in the cluster.

      and

    - You want to retain the current configuration of the rest of the cluster manager software on all the other cluster components.

- `ha=2`, which specifies that you want to configure HA ICE leader node 2, which consists of two individual nodes. Use the `ha=` field this way in the following situation:

    - Your plan is to install (or reinstall) cluster manager software on HA ICE leader node 2 in the cluster.

      and

    - You want to retain the current configuration of the rest of the cluster manager software on all the other cluster components.

3. Proceed to the following:

   Updating the software repository

Updating the software repository

This topic explains how to update the software in the repositories that you created with the cluster configuration tool. This software update involves the following tasks:

- Synchronizing the software repository
- Installing software updates
- Cloning images

The following procedure assumes that the cluster has a connection to the Internet. To perform this procedure on a secure cluster, modify this procedure. For a secure system, obtain the software updates from the HPE customer portal manually.

Procedure

1. Through an `ssh` connection, log into the admin node as the root user.
2. Retrieve the updated packages from the HPE customer portal and the operating system vendor.

   The cluster manager release notes describe how to configure local mirrors. The following Knowledge Base article also discusses this process:

   https://support.hpe.com/hpsc/doc/public/display?docId=emr_na-a00049010en_us

   For RHEL-based systems, make sure that the system is subscribed for operating system updates.

   This step requires that the system be connected to the Internet. Contact your technical support representative if this update method is not acceptable for your site.
3. Enter the `cm image show` command to retrieve the image names.

   For example:

   ```
   # cm image show
   rhel8.1
   ice-rhel8.1
   lead-rhel8.1
   ```

   NOTE: The `cm image show` command does not display scalable unit (SU) leader node image names at this point in the installation. A later procedure explains how to create SU leader node images. Even if you plan to configure SU leader nodes, use the instructions in the following steps to update the images for the non-ICE compute nodes.
4. (Optional) Back up the existing images to the cluster manager version control system.

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

   Enter the following command:

   ```
   cm image copy -s src_image_name -i image -r revision
   ```

   For *src_image_name*, specify the name of the source image. For example: `sles15sp1`.

   For *image*, specify a file name for the copied file (the clone). For example: `copy-sles15sp1`.

   For *revision*, specify a revision number

   Example 1: The following commands create backup copies of the current installation images:

   ```
   # cm image copy -s ice-compute-rhel8.1 -i ice-compute-rhel8.1.backup -r2
   # cm image copy -s rhel8.1 -i rhel8.1.backup
   # cm image copy -s lead-rhel8.1 -i lead-rhel8.1.backup
   ```

Example 2: In this example, the commands create backup copies of the current installation images and tag the backup copies as source-controlled copies. The commands assume that there are multiple versions of the source image that exist at this time. The commands copy revision 2 of the source image to the backup.

```
# cm image copy -s ice-compute-rhel8.1 -i ice-compute-rhel8.1.backup -r 2
# cm image copy -s rhel8.1 -i rhel8.1.backup -r 2
# cm image copy -s lead-rhel8.1 -i lead-rhel8.1.backup -r 2
```

5. Update the operating system software in the node image.

   Use the `cm image update` command.

   For example, to install the packages shown in Step **3**, enter the following commands:

```
# cm image update -i ice-compute-rhel7.7
# cm image update -i rhel7.7
# cm image update -i lead-rhel7.7
```

6. (Conditional) Create images for non-ICE compute nodes that do not match the architecture or operating system of the admin node.

   Complete this step as needed. For information about managing software images, see the following:

   HPE Performance Cluster Manager Administration Guide

   Obtain the operating system DVDs you need from the operating system vendor. Use the following cluster manager DVDs as needed:

   - HPE Performance Cluster Manager 1.3.1 Repository Setup DVD RHEL 8 (aarch64)
   - HPE Performance Cluster Manager 1.3.1 Repository Setup DVD SLES15 SP1 (aarch64)
   - HPE Performance Cluster Manager 1.3.1 Repository Setup DVD SLES12 SP4 (aarch64)

7. Proceed to one of the following:

   - To configure scalable unit (SU) leader nodes, proceed to the following:

     Creating scalable unit (SU) leader node images and configuration files

   - To configure a cluster without SU leader nodes, proceed to the following:

     Verifying the cluster definition file

Creating scalable unit (SU) leader node images and configuration files

Complete the procedures in this chapter if you want to configure SU leader nodes. You can configure SU leader nodes if your cluster is an HPE Apollo cluster.

NOTE: You cannot configure SU leader nodes for an HPE SGI 8600 cluster. These clusters include hardware-specific leader nodes.

If you are reinstalling a cluster system with SU leader nodes, configure the same number of SU leader nodes that the cluster had when you took delivery from the HPE factory. It is possible to change the number of SU leader nodes in a cluster. If you want to change the number of SU leader nodes, consult your sales representative. When you purchased your cluster, your HPE sales representative helped you plan your configuration. The HPE sales representative you worked with can help you to plan a different configuration.

The following topics explain SU leader node files and how to begin the SU leader node configuration process:

- **Files used when configuring scalable unit (SU) leader nodes**
- **Creating a scalable unit (SU) leader node image and editing the `su-leader-setup.conf` file**

Files used when configuring scalable unit (SU) leader nodes

You use the following files when you configure SU leader nodes for a cluster:

- `/opt/clmgr/etc/su-leader-setup.conf`

  This is the SU leader node setup file. This file specifies the number of SU leader nodes in the cluster.

- `/opt/clmgr/etc/su-leader-nodes.lst`

  This is the SU leader node list file. This file describes the SU leader node hostnames, IP addresses, and their shared LUN.

  You can obtain a copy of the original version of this file from HPE if you need it in the future.

- The cluster definition file. This is the main configuration file. When you configure SU leader nodes, you need to split the file into the following files:

  - The switch and SU leader node definition file
  - The compute node definition file, which defines the nodes you can deploy for user services

  A later procedure describes how to split the cluster definition file. When you configure a cluster with SU leader nodes, you run the `discover` command twice, once on each file.

  After you configure the cluster, if you dump the cluster definition file, the cluster manager dumps one file, not two. You can use a text editor to create two different files if you need to reinstall the cluster.

  You can obtain a copy of the original version of this file from HPE if you need it in the future.

Creating a scalable unit (SU) leader node image and editing the su-leader-setup.conf file

The procedure in this topic starts the SU leader node configuration process. The cluster manager requires an x86_64 architecture for SU leader nodes.

Procedure

1. Use the `cm image create` command to create an SU leader node image.

   The format for this command is as follows:

   ```
   cm image create -i new_image_name -l path_to_rpmlist
   ```

   Examples:

   For RHEL 8.1:

   ```
   # cm image create -i su-rhel8.1 -l /opt/clmgr/image/rpmlists/generated/generated-rhel8.1.rpmlist
   ```

   For RHEL 8.0:

   ```
   # cm image create -i su-rhel8 -l /opt/clmgr/image/rpmlists/generated/generated-rhel8.rpmlist
   ```

   For RHEL 7.7:

   ```
   # cm image create -i su-rhel7.7 -l /opt/clmgr/image/rpmlists/generated/generated-rhel7.7.rpmlist
   ```

   For SLES 15 SP1:

   ```
   # cm image create -i su-sles15sp1 -l /opt/clmgr/image/rpmlists/generated/generated-sles15sp1.rpmlist
   ```

   For SLES 12 SP5:

   ```
   # cm image create -i su-sles12sp5 -l /opt/clmgr/image/rpmlists/generated/generated-sles12sp5.rpmlist
   ```

   For SLES 12 SP4:

   ```
   # cm image create -i su-sles12sp4 -l /opt/clmgr/image/rpmlists/generated/generated-sles12sp4.rpmlist
   ```

2. Add operating system packages to the SU leader node image.

   These packages include support for the Gluster file system and for the CTDB database. The commands for this step are specific to your operating system.

   - Enter the following command for RHEL 8.1:

     ```
     # cm image dnf -i su-rhel8.1 install su-leader-collection
     ```

   - Enter the following command for RHEL 8.0:

     ```
     # cm image dnf -i su-rhel8 install su-leader-collection
     ```

   - Enter the following command for RHEL 7.7:

     ```
     # cm image yum -i su-rhel7.7 install su-leader-collection
     ```

   - Enter the following command for SLES 15 SP1:

     ```
     # cm image zypper -i su-sles15sp1 install su-leader-collection
     ```

- Enter the following command for SLES 12 SP5:

  ```
  # cm image zypper -i su-sles12sp5 install su-leader-collection
  ```

- Enter the following command for SLES 12 SP4:

  ```
  # cm image zypper -i su-sles12sp4 install su-leader-collection
  ```

3. Proceed to the following:

   Verifying the cluster definition file

Verifying the cluster definition file

The `discover` command adds a component into the cluster database and completes any required configuration tasks for a given component. A **cluster definition file** contains the following:

- A list of cluster components
- Component-specific characteristics that need to be specified

Before you run the `discover` command, complete the following procedure to verify that the cluster definition file is formatted in the correct manner.

Procedure

1. Retrieve a copy of the cluster definition file.

For clusters that are configured with at least one working slot, enter the following command to generate a cluster definition file:

```
discover --show-configfile --skip-examples > file_name
```

For *file_name*, specify any file name. You can write the cluster definition file to any directory.

The following are additional notes regarding the cluster definition file:

- The `discover` command shown in this step writes one cluster definition file to *file_name*. This file lists all the cluster components.
- You might need to split the cluster definition file into two or three cluster definition files. If you have two or three definition files, each file can contain entries only for a given type of component.
- If necessary, you can obtain the cluster definition file used in the manufacturing process from your technical support representative.
- The HPE factory uses a default cluster definition file in the following location:

  ```
  /var/tmp/mfgconfigfile
  ```

2. (Conditional) Split the cluster definition file into additional files, as needed.

Complete this step as follows:

- If you have an HPE Apollo cluster that is not an HPE Apollo 9000, complete this step.

  In this case, you need to split the cluster definition file into at least two files. With one file, configure the scalable unit (SU) leader nodes and management switches. Use the second file to configure the non-ICE compute nodes.
- If you want a more granular, step-by-step approach to the order in which components are configured into the cluster, you can complete this step.

NOTE: If the cluster is an HPE Apollo 9000, do not complete this step.

The chassis management controllers (CMCs) in HPE Apollo 9000 clusters facilitate automatic compute node configuration. For these clusters, the cluster definition file contains only the information necessary to configure the leader nodes and the switches. Later in this process, you run a command to configure the non-ICE compute nodes.

The following are example cluster definition file names and content:

Example file name	Content
mgmtsw.config	Management switches only
mgmtsw_suleader.config	Management switches and SU leader nodes
compute.config	Compute nodes

The following is one method for splitting a single configuration file into multiple files:

a. Use the cp command to copy the original cluster definition file to another file.

b. Open the file(s) you just created, and search for the [discover] section. Use an editor such as vim.

c. Retain the lines that pertain to the components for which the file is named. Delete the other lines. For example, in a file for management switches, delete the lines that pertain to leader nodes and compute nodes. The file should contain only lines for management switches.

d. Review these files carefully before proceeding.

The following examples show how to create a component-specific configuration file and show the content of the file. The files show parts of cluster definition files for various components. The ellipsis (...) indicates that the lines can be longer and include more information.

Example 1. To create a cluster definition file for management switches only, enter the following:

```
# cp /var/tmp/mfgconfigfile mgmtsw.config
```

After editing, mgmtsw.config looks like this:

```
[discover]
internal_name=mgmtsw0, type=spine, ...
internal_name=mgmtsw1, type=leaf, ...
```

Example 2. To create a cluster definition file for management switches and SU leader nodes, enter the following:

```
# cp /var/tmp/mfgconfigfile mgmtsw_suleader.config
```

After editing, mgmtsw_suleader.config looks like this:

```
[discover]
internal_name=mgmtsw0, type=spine, ...
internal_name=mgmtsw1, type=leaf, ...
internal_name=service100, su_leader_role=yes, ...
internal_name=service101, su_leader_role=yes, ...
```

Example 3. To create a cluster definition file for flat compute nodes only, enter the following:

```
# cp /var/tmp/mfgconfigfile compute.config
```

After editing, compute.config looks like this:

```
[discover]
internal_name=service0, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, ...
internal_name=service1, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, ...
```

For more information, see the following:

Cluster definition file contents

Cluster definition file examples with node templates, network interface card (NIC) templates, and predictable names

3. (Conditional) Create a custom partitions configuration file.

   Complete this step only if you want to create custom partitions on one or more non-ICE compute nodes.

   The following information pertains to custom partitions on non-ICE compute nodes:

   - If the admin node is configured to use default partitions, you can create custom partitions on non-ICE compute nodes.
   - If the admin node is configured to use custom partitions, you can create custom partitions on non-ICE compute nodes that use a different partitioning scheme.
   - You can create custom partitions on any non-ICE compute node, and the partitions can be different on each non-ICE compute node. Create one custom partitioning file for each partitioning scheme that you want to impose on one or more non-ICE compute nodes.

   The following steps explain how to create one configuration file for custom partitions on one or more noon-ICE compute nodes:

   - Change to the following directory:

     ```
     /opt/clmgr/image/scripts/pre-install
     ```

   - Open file `custom_partitions.cfg`.

     This name is the default name for the custom partition configuration file, but you can rename this file as needed. You can create multiple files. If you create multiple files, you can use any names for the files.

   - Use the guidelines in the custom partition configuration file to describe the custom partitions you want to create.
   - Save and close the custom partition configuration file.
   - Open the cluster definition file(s).
   - For each non-ICE compute node that you want to configure with custom partitions, locate the setting for each node. Add the following setting, which points to the custom partition configuration file:

     ```
     custom_partitions=file.cfg
     ```

     For example, assume that node `service1` uses the partition layout specified in custom partition file `custom_partitions_new.cfg`. You could have the following specification in the cluster definition file(s):

     ```
     internal_name=service1, mgmt_bmc_net_name=head-bmc,
     mgmt_bmc_net_macs=0c:c4:7a:c0:77:fc, mgmt_net_name=head,
     mgmt_net_macs="0c:c4:7a:c0:7a:00,0c:c4:7a:c0:7a:01", hostname1=r01n02,
     rootfs=disk, transport=udpcast, redundant_mgmt_network=no,
     switch_mgmt_network=yes, dhcp_bootfile=ipxe,
     conserver_logging=yes, conserver_ondemand=no, console_device=ttyS1,
     custom_partitions=custom_partitions_new.cfg
     ```

   - Save and close the cluster definition file(s).

   For more information about custom partitioning, see the following:

   (Conditional) Configuring custom partitions on the admin node

4. Proceed to the following:

   Running the `discover` command to complete the cluster configuration

Cluster definition file contents

Hewlett Packard Enterprise recommends that you use a cluster definition file when you configure and bring up the cluster system. When you use a cluster definition file, all the cluster configuration data resides in files that are easy to maintain and easy to edit. The cluster definition file also removes uncertainly when you configure the cluster. When you use the `discover` command, specify the cluster definition file and specify the nodes and components to discover and configure.

The following table shows the types of cluster components in the cluster definition file:

Component	**Cluster types that can include the component**	**Notes**
Management switches.	All clusters	If you use management switches that HPE does not support, do not include them in the cluster definition file. Use the configuration instructions from the manufacturer.
Power distribution units (PDUs).	HPE Apollo 9000 HPE SGI 8600 SGI Rackable	The HPE Apollo 9000 clusters have PDUs. The cluster manager configures the PDUs into the cluster automatically when the `discover` command runs. Other HPE Apollo clusters do not have PDUs that work with the cluster manager. PDUs are numbered starting with `0`. For example, `pdu0`, `pdu1`, `pdu2`, and so on.
Scalable unit (SU) leader nodes.	HPE Apollo clusters	SU leader nodes are numbered starting with 1. For example, `leader1`, `leader2`, and so on.
Rack leader controller (RLC) nodes. Also called ICE leader nodes.	HPE SGI 8600	These ICE leader nodes are numbered starting with 1. For example, on HPE SGI 8600 clusters, `r1lead`, `r2lead`, and so on. The cluster manager configures the ICE compute nodes into the cluster automatically when the `discover` command runs. There is no need to include non-ICE compute nodes in a cluster definition file.
Non-ICE compute nodes.	All clusters	On most clusters, include the non-ICE compute nodes in a cluster definition file. On HPE Apollo 9000 clusters, the cluster manager configures the non-ICE compute nodes into the cluster automatically when the `discover` command runs. Do not include non-ICE compute nodes in a cluster definition file. To assign site-specific hostnames to these nodes, assign the hostnames after the cluster is configured.

In the cluster definition file, each component is defined with several variables. For example, these variables can include MAC addresses, IP addresses, component roles, hostnames, management network details, the node image assignment, and much more.

By default, HPE configures nodes with hostnames that correspond to their default number, as follows:

- Non-ICE compute nodes that belong to the general pool of computing resources are numbered starting with `0`. For example, `n0`, `n1`, and so on. For example, on HPE Apollo 9000 clusters, the factory configures these nodes with names such as `r1c1t1n2`, `r1c3t1n4`.

  Non-ICE compute nodes with services installed upon them are numbered starting with 0. For example, `service0`, `service1`. These names are the default names for non-ICE compute nodes that are under cluster manager control. The `service` part of the name distinguishes nodes that host services from nodes that are among the pool of general computing resources.

- ICE compute nodes are assigned into the general pool of computing resources. These nodes are associated with an ICE leader node and are numbered starting with 0. The numbering depends on the following:

  - The ICE leader node number
  - The chassis number within the ICE leader node

  For example, the factory numbers the first blade on ICE leader node 1, chassis 1 as `r1i0n0`. If there are eight chassis in the rack, the factory numbers the last blade on the last chassis of ICE leader node as `r1i0n7`.

  Typically, ICE compute nodes are not deployed for user services. They are numbered starting with 1. For example, the factory configures these nodes with names such as `r01n01`, `r01n02`, `r01n03`.

- Graphic processing units (GPUs) are numbered starting with 1. For example, the factory configures graphical compute nodes with names such as `r01g01`.

For information about the cluster definition file variables, enter one of the following commands:

- `# **man discover**`
- `# **discover -h**`

If you no longer have the cluster definition file for the cluster, you can obtain the original cluster definition file from the HPE factory. Another way to obtain a cluster definition file is to enter the following command and build a file from the resulting file:

```
# discover --show-configfile
```

Both the `discover` command and the `configure-cluster` command accept a cluster definition file as input. To specify a configuration file to these commands, add the `--configfile` *`file`* parameter.

Cluster definition file examples with node templates, network interface card (NIC) templates, and predictable names

Contemporary cluster definition files contain node template sections and use predictable NIC names. Use the following keywords at the start of sections in the file that pertain to node templates and NIC templates:

- `[templates]`

  The cluster manager assumes that the lines following the `[templates]` keyword define the characteristics for a specific node type.

  For example, you can define templates for the non-ICE compute nodes and the leader nodes. Templates are useful when they pertain to multiple nodes, for example, many identical non-ICE compute nodes. You can describe the nodes once, in the template section of the cluster definition file. After you run the `discover` command, you no longer need the template. The node template definitions can describe kernel names, image names, BMC authentication info, and other node characteristics.

For more information, see the node-templates(8) manpage.

- [nic_templates]

  NIC templates pertain to the NIC devices in specific nodes. Each node template can have one or more NIC templates. The NIC templates explain how to tie networks to interfaces. There can be one NIC template per network. The NIC template definitions can describe the network interfaces for the network, the network name, bonding settings, and so on.

  If you want to have a [nic_templates] section in the cluster definition file, also create a [templates] section.

  Predictable names pertain to the NICs within each node. These NIC names are the same across like hardware.

  If you have an HA admin node, the two physical admin nodes use legacy names. The HA admin node, which is a virtual machine, uses predictable names.

  InfiniBand devices do not use predictable names.

  For more information about predictable names, see the following:
  Predictable network interface card (NIC) names

By default, the cluster manager reads in templates from the following file when you run the cluster configuration tool:

```
/etc/opt/sgi/default-node-templates.conf
```

Cluster definition file example - HPE Apollo 9000 cluster

To configure an HPE Apollo 9000 cluster, use one cluster definition file. For this HPE Apollo 9000 systems, the cluster definition file defines the switches and the scalable unit (SU) leader nodes. You do not need to define the non-ICE compute nodes in a cluster definition file for the HPE Apollo 9000. The cluster manager configures the non-ICE compute nodes automatically.

Explanations for the items in bold print in are as follows:

- The [templates] section includes information that pertains to all of the SU leader nodes. This section contains one line, and that line begins with the name= field. This line defines the su-leader node type and sets the characteristics for nodes of type su-leader.

  The image= field defines the image that you want the installer to put on the SU leader node.

- The [discover] section includes lines for each SU leader node.

  The template_name= field appears in the definition lines for each SU leader node. This field identifies the node as being of type su-leader. The installer applies the characteristics defined in the [templates] section to the nodes that include template_name=su-leader.

  The hostname1= field defines the hostname for the SU leader node.

```
# File apollo9000.config
# Cluster definition file for management switches and SU leader nodes on an
HPE Apollo 9000 cluster
# /bin/bash

[templates]
name=service, console_device=ttyS0, conserver_logging=yes,
mgmt_net_name=hostmgmt2001, mgmt_bmc_net_name=hostctrl2001,
rootfs=tmpfs, transport=bt, mgmt_net_bonding_master=bond0,
dhcp_bootfile=grub2, disk_bootloader=no, mgmt_net_interfaces="eno1",
switch_mgmt_network=yes, tpm_boot=no, conserver_ondemand=no,
redundant_mgmt_network=no, predictable_net_names=yes,
card_type=iLO, baud_rate=115200, bmc_username=Administrator,
bmc_password=compaq
name=su-leader, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head,
mgmt_net_interfaces="eno5,eno6", rootfs=disk,
transport=rsync, predictable_net_names=yes, switch_mgmt_network=yes,
redundant_mgmt_network=yes, console_device=ttyS0,
architecture=x86_64, card_type=iLO, image=su-rhel8.1,
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=802.3ad,
bmc_username=admin, bmc_password=admin, baud_rate=115200

[nic_templates]
template=service, network=hostmgmt2001, bonding_master=bond0,
bonding_mode=active-backup, net_ifs="eno1"

[discover]
# admin node
internal_name=admin, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="48:df:
37:89:45:90", mgmt_bmc_net_ip=172.24.0.1,
mgmt_net_name=head, mgmt_net_macs="48:df:37:89:45:90,48:df:37:89:45:98",
mgmt_net_interfaces="eno5,eno6",
mgmt_net_ip=172.23.0.1, admin_house_interface=eno1, hostname1=viking,
rootfs=disk, transport=rsync, redundant_mgmt_network=yes,
switch_mgmt_network=yes, predictable_net_names=yes, console_device=ttyS0,
architecture=x86_64, mgmt_net_bonding_mode=802.3ad
# management switches
internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="ec:9b:8b:
60:7e:b0", redundant_mgmt_network=yes, net=head/head-bmc,
type=spine, ice=no
internal_name=mgmtsw1, mgmt_net_name=head, mgmt_net_macs="4c:ae:a3:2d:
05:80", redundant_mgmt_network=no, net=head/head-bmc,
type=leaf, ice=no
# SU leaders
internal_name=service1, mgmt_bmc_net_macs="20:67:7c:e4:f3:4c",
mgmt_net_macs="48:df:37:87:d0:80,48:df:37:87:d0:88",
hostname1=leader1, template_name=su-leader, image=su-rhel8.1
internal_name=service2, mgmt_bmc_net_macs="20:67:7c:e4:f3:1c",
mgmt_net_macs="48:df:37:87:a8:20,48:df:37:87:a8:28",
hostname1=leader2, template_name=su-leader, image=su-rhel8.1
internal_name=service3, mgmt_bmc_net_macs="20:67:7c:e4:f3:36",
mgmt_net_macs="48:df:37:87:a6:a0,48:df:37:87:a6:a8",
hostname1=leader3, template_name=su-leader, image=su-rhel8.1
# badger compute nodes (routed VLANs)

[dns]
cluster_domain=cm.chf.rdlabs.hpecorp.net
nameserver1=16.110.135.51
nameserver2=16.110.135.52

[attributes]
admin_house_interface=eno1
admin_mgmt_interfaces="eno5,eno6"
admin_mgmt_bmc_interfaces="eno5,eno6"
admin_udpcast_ttl=2
admin_udpcast_mcast_rdv_addr=239.255.255.1
admin_mgmt_bonding_mode=802.3ad
dhcp_bootfile=grub2
udpcast_max_wait=10
```

```
udpcast_rexmit_hello_interval=0
predictable_net_names=yes
udpcast_mcast_rdv_addr=224.0.0.1
copy_admin_ssh_config=yes
udpcast_max_bitrate=900m
udpcast_min_wait=10
domain_search_path=ib0.acme.net,sfo.acme.net,msp.acme.net,lux.acme.net
udpcast_min_receivers=1
mcell_network=yes
head_vlan=1
conserver_logging=yes
switch_mgmt_network=yes
redundant_mgmt_network=yes
rack_vlan_end=1100
max_rack_irus=16
blademond_scan_interval=120
rack_vlan_start=101
mcell_vlan=3
conserver_ondemand=no
monitoring_ganglia_enabled=yes
monitoring_native_enabled=yes
monitoring_nagios_enabled=yes
monitoring_kafka_elk_alerta_enabled=yes

[networks]
name=public, subnet=137.38.83.0, netmask=255.255.255.0, gateway=137.38.83.1
name=head, type=mgmt, vlan=1, subnet=172.23.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0,
netmask=255.255.0.0, rack_netmask=255.255.252.0
name=hostctrl, type=mgmt-bmc, subnet=10.171.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0
name=hostctrl2001, type=mgmt-bmc, vlan=2001, subnet=10.171.0.0,
netmask=255.255.252.0, gateway=10.171.3.254
name=hostmgmt, type=mgmt, subnet=10.170.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0
name=hostmgmt2001, type=mgmt, vlan=2001, subnet=10.170.0.0,
netmask=255.255.252.0, gateway=10.170.3.254
name=ib-0, type=ib, vlan=1, subnet=10.148.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0, gateway=172.23.255.254
name=ib-1, type=ib, vlan=1, subnet=10.149.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0, gateway=172.23.255.254
name=gbe, type=lead-mgmt, vlan=1, subnet=10.159.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0, gateway=128.162.243.1
name=bmc, type=lead-bmc, subnet=10.160.0.0, netmask=255.255.0.0,
rack_netmask=255.255.252.0
name=mcell-net, type=cooling, subnet=172.26.0.0, netmask=255.255.0.0
name=ha-net, type=ha, vlan=1, subnet=192.168.161.0, netmask=255.255.255.0,
rack_netmask=255.255.252.0, gateway=172.23.255.254
name=ha-corosync-mcast0, type=ha, subnet=226.95.0.0, netmask=255.255.255.0,
rack_netmask=255.255.252.0
name=ha-corosync-mcast1, type=ha, vlan=1, subnet=226.96.0.0,
netmask=255.255.255.0, rack_netmask=255.255.252.0, gateway=172.23.255.254

[images]
image_types=default
```

Cluster definition file example - HPE Apollo cluster with scalable unit (SU) leader nodes

To configure an HPE Apollo cluster with SU leader nodes, use two cluster definition files:

- The switch and SU leader definition file.
- The non-ICE compute node definition file. This file defines the general compute nodes and the compute nodes that you want to configure with user services

NOTE: The information in this topic does not apply to HPE Apollo 9000 clusters. For information about the cluster definition file for HPE Apollo 9000 clusters, see the following:

Cluster definition file example - HPE Apollo 9000 cluster

The following is an example file for the switches and the SU leader nodes. Explanations for the items in bold print in `mgmtsw_suleader.config` are as follows:

- The `[templates]` section includes information that pertains to all of the SU leader nodes. This section contains one line, and that line begins with the `name=` field. This line defines the `su-leader` node type and sets the characteristics for nodes of type `su-leader`.

  The `image=` field defines the image that you want the installer to put on the SU leader node.

- The `[discover]` section includes lines for each SU leader node.

  The `template_name=` field appears in the definition lines for each SU leader node. This field identifies the node as being of type `su-leader`. The installer applies the characteristics defined in the `[templates]` section to the nodes that include `template_name=su-leader`.

  The `hostname1=` field defines the hostname for the SU leader node.

```
# File mgmtsw_suleader.config
# Cluster definition file for management switches and SU leader nodes on an HPE Apollo cluster
[templates]
name=su-leader, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1,eno2",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=802.3ad, redundant_mgmt_network=yes,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=disk, card_type=iLO, baud_rate=115200,
force_disk="/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:0", su_leader_role=yes

[nic_templates]
template=su-leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="eno1,eno2"
template=su-leader, network=head-bmc, net_ifs="bmc0"
template=su-leader, network=ib-0, net_ifs="ib0"
template=su-leader, network=ib-1, net_ifs="ib1"

[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine, mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.1
internal_name=mgmtsw2, mgmt_net_macs="40:b9:3c:a6:6a:a2", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.2
internal_name=service1, hostname1=leader1, mgmt_bmc_net_macs="20:67:7c:e4:8a:8a",
mgmt_net_macs="00:0f:53:21:98:30,00:0f:53:21:98:31", mgmt_net_ip=172.23.10.1,
mgmt_bmc_net_ip=172.24.10.1, template_name=su-leader
internal_name=service2, hostname1=leader2, mgmt_bmc_net_macs="20:67:7c:e4:9a:ba",
mgmt_net_macs="00:0f:53:21:98:90,00:0f:53:21:98:91", mgmt_net_ip=172.23.10.2,
mgmt_bmc_net_ip=172.24.10.2, template_name=su-leader
internal_name=service3, hostname1=leader3, mgmt_bmc_net_macs="20:67:7c:e4:8a:7a",
mgmt_net_macs="00:0f:53:3c:e0:a0,00:0f:53:3c:e0:a1", mgmt_net_ip=172.23.10.3,
mgmt_bmc_net_ip=172.24.10.3, template_name=su-leader
```

The following is an example file for the non-ICE compute nodes attached to the SU leader nodes:

```
# File su_compute.config
# Cluster definition file for compute nodes that utilize an SU leader on an HPE Apollo cluster
[templates]
name=su-compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1,eno2",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, redundant_mgmt_network=yes,
switch_mgmt_network=yes, transport=rsync, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=nfs, card_type=iLO, baud_rate=115200, bmc_username=Administrator, bmc_password=compaq

[nic_templates]
template=su-compute, network=head, bonding_master=bond0, bonding_mode=active-backup, net_ifs="eno1,eno2"
template=su-compute, network=head-bmc, net_ifs="bmc0"
template=su-compute, network=ib-0, net_ifs="ib0"
template=su-compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=service101, mgmt_bmc_net_macs="20:67:7c:e4:9a:10",
mgmt_net_macs="00:0f:53:21:98:11,00:0f:53:21:98:12", mgmt_bmc_net_ip=172.24.1.1, mgmt_net_ip=172.23.1.1,
template_name=su-compute, su_leader=172.23.255.241
internal_name=service102, mgmt_bmc_net_macs="20:67:7c:e4:9a:21",
mgmt_net_macs="00:0f:53:21:98:22,00:0f:53:21:98:23", mgmt_bmc_net_ip=172.24.1.2, mgmt_net_ip=172.23.1.2,
template_name=su-compute, su_leader=172.23.255.242
internal_name=service103, mgmt_bmc_net_macs="20:67:7c:e4:9a:32",
mgmt_net_macs="00:0f:53:21:98:33,00:0f:53:21:98:34", mgmt_bmc_net_ip=172.24.1.3, mgmt_net_ip=172.23.1.3,
template_name=su-compute, su_leader=172.23.255.243
internal_name=service201, mgmt_bmc_net_macs="20:67:7c:e4:9a:43",
mgmt_net_macs="00:0f:53:21:98:44,00:0f:53:21:98:45", mgmt_bmc_net_ip=172.24.2.1, mgmt_net_ip=172.23.2.1,
template_name=su-compute, su_leader=172.23.255.241
internal_name=service202, mgmt_bmc_net_macs="20:67:7c:e4:9a:54",
mgmt_net_macs="00:0f:53:21:98:55,00:0f:53:21:98:56", mgmt_bmc_net_ip=172.24.2.2, mgmt_net_ip=172.23.2.2,
template_name=su-compute, su_leader=172.23.255.242
internal_name=service203, mgmt_bmc_net_macs="20:67:7c:e4:9a:65",
mgmt_net_macs="00:0f:53:21:98:66,00:0f:53:21:98:67", mgmt_bmc_net_ip=172.24.2.3, mgmt_net_ip=172.23.2.3,
template_name=su-compute, su_leader=172.23.255.243
```

Cluster definition file example - Cluster with HPE Apollo Moonshot system cartridges

The following procedure explains how to configure HPE Moonshot system cartridges into a cluster.

Procedure

1. Obtain the IP address of the HPE Moonshot chassis.

   If the chassis is configured with a static IP address, connect to the chassis console and determine the iLOCM IP address.

   If the chassis is configured to use DHCP, complete the following steps:

   a. Power on (plug in) the chassis.

      You do not need to power-on the individual cartridges.

      For cabling information, see the following:

      HPE Moonshot 1500 Chassis Setup and Installation Guide

   b. Log into the admin node as the root user.

   c. Monitor the `/var/log/messages` file.

      Use a command such as `tail -f`.

   d. Wait for an entry that shows the `DHCPDISCOVER` line that includes the MAC address of the iLOCM, and observe the chassis IP address in the lines that follow.

For example:

```
May 19 15:36:38 cmutay1 dhcpd: DHCPDISCOVER from 9c:b6:54:8a:28:72 via eth0
May 19 15:36:38 cmutay1 dhcpd: DHCPOFFER on 10.117.23.6 to 9c:b6:54:8a:28:72 via eth0
May 19 15:36:42 cmutay1 dhcpd: DHCPREQUEST for 10.117.23.6 (10.117.20.74) from 9c:b6:54:8a:28:72 via eth0
May 19 15:36:42 cmutay1 dhcpd: DHCPACK on 10.117.23.6 to 9c:b6:54:8a:28:72 via eth0
```

The IP address is `10.117.23.6`.

2. From the admin node, use the `cm_scan_moonshot` command to generate information that you can include in the cluster definition file.

   This step generates node definitions for all the cartridges in the HPE Moonshot system chassis.

   The `cm_scan_moonshot` command has several parameters. The following command line shows the basic parameters needed to generate information for the cluster definition file:

   ```
   cm_scan_moonshot -L ilocm_ip(s) -G ["string"] -n name_syntax -o outfile
   ```

Variable	Information
ilocm_ip(s)	The IP address of one or more iLOCMs, that is the iLO chassis managers. If you specify more than one IP address, use a comma (`,`) to separate each address.
string	Optional. A string of node information that you want the cluster manager to write to the output file. Enclose the string in quotation marks (`"  "`).
name_syntax	A pattern for the generated node names. You can include the wildcard characters that this command supports. For a list of these characters, enter the following: `# cm_scan_moonshot -h` For example, if you specify `-n node%2i`, the command generates node names that start with `node` and have a 2-integer suffix. That is, in the cluster definition file, the nodes are numbered as `node01`, `node02`, `node03`, ... `node99`.
outfile	An output file name.

For example:

- Example 1. Assume that you have one chassis, and you want to configure the 10 cartridges in that chassis into an HPE Apollo cluster. Enter the following command to generate node definitions:

  ```
  # cm_scan_moonshot -L 172.24.5.5 \
  -G "tpm_boot=no, predictable_net_names=yes, force_disk=/dev/sda, destroy_disk_label=yes" \
  -n node%2i -o /tmp/moonshot.txt
  ```

  This command scans the HPE Moonshot system chassis at `172.24.5.5` and generates a file called `moonshot.txt`. The file contains a series of 10 node definitions suitable for appending to a cluster definition file and is as follows:

  ```
  internal_name=service01, hostname1=node01, mgmt_net_macs=38:ea:a7:0f:48:08, mgmt_bmc_net_macs=38:ea:a7:0f:66:fe,
  mgmt_bmc_net_ip=172.24.5.5, card_type=ILOCM, architecture=x86_64, console_device=ttyS0, baud_rate=9600,
  bmc_username=admin, bmc_password=admin123, tpm_boot=no, predictable_net_names=yes, force_disk=/dev/sda,
  destroy_disk_label=yes
  internal_name=service02, hostname1=node02, mgmt_net_macs=38:ea:a7:0f:3d:b6, mgmt_bmc_net_macs=38:ea:a7:0f:66:fe,
  mgmt_bmc_net_ip=172.24.5.5, card_type=ILOCM, architecture=x86_64, console_device=ttyS0, baud_rate=9600,
  bmc_username=admin, bmc_password=admin123, tpm_boot=no, predictable_net_names=yes, force_disk=/dev/sda,
  ```

```
destroy_disk_label=yes
.
.
.
```

The -G parameter appends the additional configuration attributes, in a comma-separated list, to the lines for each compute node.

- Example 2. Assume that you have two chassis and that you want to generate node definitions in the output file that include the configuration attribute `predictable_net_names=yes`. Enter the following command:

```
# cm_scan_moonshot -L 172.24.5.4,172.24.5.5 \
-G "predictable_net_names=yes" -n node%3i -o /tmp/moonshot.txt
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.
45 nodes scanned from ILOCM 172.24.5.4
45 nodes scanned from ILOCM 172.24.5.5

Scanning complete. 90 node(s) written to file /opt/clmgr/tmp/tmp_scan_file-20077
Final scan results written to file: /tmp/moonshot.txt
```

3. Open the output file and the cluster definition file in a text editor.

4. Find the `[discover]` section in the cluster definition file, and add the lines from the output file at the end of the `[discover]` section.

The following shows a cluster definition file that contains lines for the HPE Apollo Moonshot system cartridges:

```
[templates]
.
.
.
[discover]
internal_name=service01, hostname1=node01, mgmt_net_macs=38:ea:a7:0f:48:08, mgmt_bmc_net_macs=38:ea:a7:0f:66:fe,
mgmt_bmc_net_ip=172.24.5.5, card_type=ILOCM, architecture=x86_64, console_device=ttyS0, baud_rate=9600,
bmc_username=admin, bmc_password=admin123, tpm_boot=no, predictable_net_names=yes, force_disk=/dev/sda,
destroy_disk_label=yes
internal_name=service02, hostname1=node02, mgmt_net_macs=38:ea:a7:0f:3d:b6, mgmt_bmc_net_macs=38:ea:a7:0f:66:fe,
mgmt_bmc_net_ip=172.24.5.5, card_type=ILOCM, architecture=x86_64, console_device=ttyS0, baud_rate=9600,
bmc_username=admin, bmc_password=admin123, tpm_boot=no, predictable_net_names=yes, force_disk=/dev/sda,
destroy_disk_label=yes
.
.
.
```

5. Use the `fastdiscover` command in the following format to configure the cartridges into the cluster:

```
fastdiscover --allow-duplicate-macs-and-ips config.file
```

For *config.file*, specify the name of the cluster definition file you edited in this procedure.

6. Use the following command to scan the chassis:

```
cm_scan_moonshot -L ilocm_ip(s)
```

For *ilocm_ip(s)*, specify the same IP address(es) that you specified in the following step:

Step **2**

These IP address(es) are for one or more iLOCMs, that is the iLO chassis managers. If you specify more than one IP address, use a comma (**,**) to separate each address.

When you use the `cm_scan_moonshot` command in this format, the command updates the cluster database with cartridge and node location information. This command is essential for proper power operations.

7. Enter the following commands:

```
# update-configs --node admin --su-leaders
# cm node refresh netboot -n "*"
```

8. Use the `cm node provision` command to provision each node with an image.

Cluster definition file example - Template that defines the characteristics for an HA ICE leader node

The template in this topic defines the characteristics of an HA ICE leader node and assumes that you want the cluster to use the default HA network. This ICE leader node template is useful for the cluster because this cluster has two HA ICE leader nodes.

The cluster includes only ICE leader nodes and ICE compute nodes. The `[images]` section of this file specifies to create only the `ice` and `lead` images. The other image type is `default`, which applies to non-ICE compute nodes. In this cluster, there are no non-ICE compute service nodes, so the `default` image type is omitted from the `[images]` section. The `[images]` section is not required in the cluster definition file, but you can use the `[images]` section to specify only the images that a cluster requires.

```
# File ha_leader.config
# Cluster definition file for HA Rack Leader Controllers (RLCs) on an HPE SGI 8600 ICE cluster
[templates]
# HA RLC template
name=ha-leader, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1,eno2",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=802.3ad, redundant_mgmt_network=yes,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2,
disk_bootloader=no, predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no,
conserver_logging=yes, rootfs=disk, card_type=ipmi, baud_rate=115200, bmc_username=admin,
bmc_password=admin, ha=all, image=lead-sles12sp5-ha, kernel=4.12.14-94.41-default
# ICE compute node template
name=ha-ice, console_device=ttyS1, rootfs=tmpfs, transport=udpcast, bmc_username=admin,
bmc_password=admin, baud_rate=115200
# Non-ICE compute node template
name=compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, redundant_mgmt_network=yes,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=disk, card_type=ipmi, baud_rate=115200, bmc_username=admin, bmc_password=admin

[nic_templates]
template=ha-leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="eno1,eno2"
template=ha-leader, network=head-bmc, net_ifs="bmc0"
template=ha-leader, network=ib-0, net_ifs="ib0"
template=ha-leader, network=ib-1, net_ifs="ib1"
# below is the interface used for RLC-HA RLC<->RLC communication
template=ha-leader, network=ha-net, net_ifs="eno4"
template=ha-ice, network=gbe, net_ifs="eno1"
template=ha-ice, network=bmc, net_ifs="bmc0"
template=ha-ice, network=ib-0, net_ifs="ib0"
template=ha-ice, network=ib-1, net_ifs="ib1"
template=compute, network=head, bonding_master=bond0, bonding_mode=active-backup, net_ifs="eno1"
template=compute, network=head-bmc, net_ifs="bmc0"
template=compute, network=ib-0, net_ifs="ib0"
template=compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=yes, type=spine, mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=yes, type=leaf, mgmt_net_ip=172.23.100.1
internal_name=mgmtsw2, mgmt_net_macs="40:b9:3c:a6:6a:a2", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=yes, type=leaf, mgmt_net_ip=172.23.100.2
internal_name=r1lead1, mgmt_bmc_net_macs="20:67:7c:e4:8a:8a",
mgmt_net_macs="00:0f:53:21:98:30,00:0f:53:21:98:31", mgmt_net_ip=172.23.10.1,
mgmt_bmc_net_ip=172.24.10.1, template_name=ha-leader
```

```
internal_name=r1lead2, mgmt_bmc_net_macs="20:67:7c:e4:8a:a3",
mgmt_net_macs="00:0f:53:21:98:a4,00:0f:53:21:98:a5", mgmt_net_ip=172.23.10.2,
mgmt_bmc_net_ip=172.24.10.2, template_name=ha-leader
internal_name=service0, mgmt_bmc_net_macs="20:67:7c:e4:9a:ba", mgmt_net_macs="00:0f:53:21:98:bb",
template_name=compute
internal_name=service1, mgmt_bmc_net_macs="20:67:7c:e4:9a:a2", mgmt_net_macs="00:0f:53:21:98:a3",
template_name=compute
internal_name=service2, mgmt_bmc_net_macs="20:67:7c:e4:9a:32", mgmt_net_macs="00:0f:53:21:98:33",
template_name=compute
internal_name=service3, mgmt_bmc_net_macs="20:67:7c:e4:9a:87", mgmt_net_macs="00:0f:53:21:98:88",
template_name=compute
.
.
.
```

Cluster definition file example - Cluster with 100 non-ICE compute nodes and no leader nodes

This example cluster definition file is for a cluster with 100 non-ICE compute nodes and no leader nodes. For simplicity, the example file shows only two non-ICE compute nodes and the management switches. The following information highlights some characteristics of this cluster:

- The information in the internal_name field defines the role for each of the two non-ICE compute nodes in this cluster. The content of the internal_name field and the hostname1 field can be identical. In other words, you can use the hostname of the node as its internal_name.

  The content of the internal_name field for each non-ICE compute node is service*n*, where *n* is a number from 1 through 101.

- The hostname1 field defines the hostname that users must specify when they want to log into a node. The hostname1 field contains the text that appears in the output for most cluster manager commands.

- The cluster definition file specifies a multicast installation that uses udpcast transport for the non-ICE compute service nodes. The non-ICE compute service nodes are service1 and service101.

- The top-level switch, mgmtsw0, is defined as spine switch. This switch is always connected to the admin node. Switch mgmtsw1 is defined as a leaf switch. Switch mgmtsw1 is connected to the spine switch, mgmtsw0.

- The definition for both switches includes ice=no because this cluster does not have leader nodes or ICE compute nodes.

- The [images] section of this file specifies to create only the default image type, which is the image type for non-ICE compute nodes. By default, the installer also creates ice and lead images, but this cluster does not need those image types. The [images] section is not required in the cluster definition file, but you can use the [images] section to specify only the images that a cluster requires.

The file is as follows:

```
# File compute.config
# Cluster definition file for regular compute nodes
[templates]
name=compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup,
redundant_mgmt_network=no, switch_mgmt_network=yes, transport=udpcast, tpm_boot=no,
dhcp_bootfile=grub2, disk_bootloader=no, predictable_net_names=yes, console_device=ttyS0,
conserver_ondemand=no, conserver_logging=yes, rootfs=disk, card_type=ipmi,
baud_rate=115200, bmc_username=admin, bmc_password=admin

[nic_templates]
template=compute, network=head, bonding_master=bond0, bonding_mode=active-backup,
net_ifs="eno1"
template=compute, network=head-bmc, net_ifs="bmc0"
```

```
template=compute, network=ib-0, net_ifs="ib0"
template=compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine,
mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.1
internal_name=service1, mgmt_bmc_net_macs="20:67:7c:e4:9a:12",
mgmt_net_macs="00:0f:53:21:98:13", template_name=compute
internal_name=service2, mgmt_bmc_net_macs="20:67:7c:e4:9a:23",
mgmt_net_macs="00:0f:53:21:98:24", template_name=compute
internal_name=service3, mgmt_bmc_net_macs="20:67:7c:e4:9a:34",
mgmt_net_macs="00:0f:53:21:98:35", template_name=compute
internal_name=service4, mgmt_bmc_net_macs="20:67:7c:e4:9a:45",
mgmt_net_macs="00:0f:53:21:98:46", template_name=compute
internal_name=service5, mgmt_bmc_net_macs="20:67:7c:e4:9a:56",
mgmt_net_macs="00:0f:53:21:98:57", template_name=compute
internal_name=service6, mgmt_bmc_net_macs="20:67:7c:e4:9a:67",
mgmt_net_macs="00:0f:53:21:98:68", template_name=compute
.
.
.
```

Cluster definition file example - Cluster with one ICE leader node, one ICE compute rack, and several non-ICE compute nodes

This example cluster definition file is for a cluster that includes the following:

- One HPE SGI 8600 ICE leader node
- One HPE SGI ICE compute rack
- Several non-ICE compute nodes

The following information highlights some characteristics of this cluster:

- This file uses templates across identical hardware types. All leader nodes in the cluster use the `leader` template. All non-ICE compute nodes use the `compute` template.
- The `leader` template specifies the image and kernel to use when a leader nodes is booted and installed
- The IP addresses for the management switches are pre-defined by the `mgmt_net_ip=` variable in the cluster definition file.

The cluster definition file is as follows:

```
# File rlc.config
# Cluster definition file for 1 ICE leader node and its ICE compute rack + several non-ice compute nodes
[templates]
# RLC template
name=leader, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1,eno2", mgmt_net_bonding_master=bond0,
mgmt_net_bonding_mode=802.3ad, redundant_mgmt_network=yes, switch_mgmt_network=yes, transport=udpcast, tpm_boot=no,
dhcp_bootfile=grub2, disk_bootloader=no, predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no,
conserver_logging=yes, rootfs=disk, card_type=ipmi, baud_rate=115200, bmc_username=admin, bmc_password=admin,
image=lead-sles15sp1, kernel=4.12.14-94.41-default, ice_template_name=ice-compute
# ICE compute node template
name=ice-compute, console_device=ttyS1, rootfs=tmpfs, transport=udpcast, bmc_username=admin, bmc_password=admin,
baud_rate=115200
# non-ICE compute node template
name=compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1", mgmt_net_bonding_master=bond0,
mgmt_net_bonding_mode=active-backup, redundant_mgmt_network=no, switch_mgmt_network=yes, transport=udpcast, tpm_boot=no,
dhcp_bootfile=grub2, disk_bootloader=no, predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no,
conserver_logging=yes, rootfs=disk, card_type=ipmi, baud_rate=115200, bmc_username=admin, bmc_password=admin

[nic_templates]
template=leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="eno1,eno2"
```

```
template=leader, network=head-bmc, net_ifs="bmc0"
template=leader, network=ib-0, net_ifs="ib0"
template=leader, network=ib-1, net_ifs="ib1"
template=compute, network=head, bonding_master=bond0, bonding_mode=active-backup, net_ifs="eno1"
template=compute, network=head-bmc, net_ifs="bmc0"
template=compute, network=ib-0, net_ifs="ib0"
template=compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head, redundant_mgmt_network=yes, net=head/head-bmc,
ice=yes, type=spine, mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head, redundant_mgmt_network=yes, net=head/head-bmc,
ice=yes, type=leaf, mgmt_net_ip=172.23.100.1
internal_name=mgmtsw2, mgmt_net_macs="40:b9:3c:a6:6a:a2", mgmt_net_name=head, redundant_mgmt_network=yes, net=head/head-bmc,
ice=yes, type=leaf, mgmt_net_ip=172.23.100.2
internal_name=r1lead, mgmt_bmc_net_macs="20:67:7c:e4:8a:8a", mgmt_net_macs="00:0f:53:21:98:30,00:0f:53:21:98:31",
mgmt_net_ip=172.23.10.1, mgmt_bmc_net_ip=172.24.10.1, template_name=leader
internal_name=service0, mgmt_bmc_net_macs="20:67:7c:e4:9a:ba", mgmt_net_macs="00:0f:53:21:98:bb", template_name=compute
internal_name=service1, mgmt_bmc_net_macs="20:67:7c:e4:9a:a2", mgmt_net_macs="00:0f:53:21:98:a3", template_name=compute
internal_name=service2, mgmt_bmc_net_macs="20:67:7c:e4:9a:32", mgmt_net_macs="00:0f:53:21:98:33", template_name=compute
internal_name=service3, mgmt_bmc_net_macs="20:67:7c:e4:9a:87", mgmt_net_macs="00:0f:53:21:98:88", template_name=compute

[dns]
cluster_domain=cm.example.domain.net
nameserver1=10.10.10.253
nameserver2=10.10.10.254
.
.
.
```

Cluster definition file example - Virtual admin node on an HA admin cluster with leader nodes

The example in this topic is a cluster definition file fragment that assigns an IP address to the storage unit. When the storage unit has an IP address, the virtual admin node can access the storage unit whenever the need arises. In addition, the file assigns IP addresses to the physical admin nodes. The presence of these IP addresses enables access to the physical admin nodes from the virtual admin node.

The file fragment is as follows:

```
# File generic_components.config
# Cluster definition file for components in the cluster that only need an IP address
[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine, mgmt_net_ip=172.23.255.254
internal_name=service50, mgmt_net_name=head, mgmt_net_macs="00:0f:45:ac:93:13",
hostname1=is5110a, discover_skip_switchconfig=yes, generic
internal_name=service51, mgmt_net_name=head, mgmt_net_macs="00:0f:45:ac:93:aa",
hostname1=is5110b, discover_skip_switchconfig=yes, generic
internal_name=service52, mgmt_net_name=head, mgmt_net_macs="00:02:aa:ac:9a:ff",
hostname1=genericnode1, discover_skip_switchconfig=yes, generic
internal_name=service53, mgmt_net_name=head, mgmt_net_macs="00:ca:31:a3:9c:b9",
hostname1=othernode1, discover_skip_switchconfig=yes, other
```

In this example, notice that the storage unit is configured.

Cluster definition file example - Configuring power distribution units (PDUs)

PDUs distribute AC power to the cluster components. PDUs are optional.

NOTE: On HPE SGI 8600 clusters, PDUs attach to the chassis management controllers (CMCs). If the cluster was configured at the factory with PDUs, you do not need to address PDUs in the installation or configuration process. The cluster manager automatically configures the PDUs for you on HPE SGI 8600 clusters.

The cluster manager does not support PDUs on SGI ICE X systems.

On clusters with server racks other than HPE Apollo 9000 racks or HPE SGI 8600 racks, the PDUs reside inside each rack. Configure PDUs explicitly by making sure that they are specified in the cluster definition file when the `discover` command runs.

For example, assume that you have a cluster with server racks other than HPE Apollo 9000 racks or HPE SGI 8600 racks. Include a line such as the following in the cluster definition file to configure the PDU numbered pdu0:

```
internal_name=pdu0, mgmt_bmc_net_name=head-bmc,
geolocation="cold isle 4 rack 1 B power",
mgmt_bmc_net_macs=99:99:99:99:99:99,
hostname1=testpdu0, pdu_protocol="snmp/mypassword"
```

Observe the following in the settings for pdu0:

- Specify the network upon which the PDU resides. In this case, it is head-bmc, which specifies the head BMC network.
- You can specify a geolocation setting. To add a text string that points to the physical location of a PDU, use the geolocation= parameter. For example:

  - hot isle 3 rack1 A power
  - cold isle 4 rack 1 B power

  The text string can include spaces and special characters. If you include spaces, enclose the string in quotation marks (").

  If you have multiple PDUs, multiple clusters, or multiple racks, this setting can be helpful. The geolocation setting is optional.
- The pdu_protocol= parameter lets you specify a protocol.

  On a cluster without ICE leader nodes, the protocol is SNMP. Specify pdu_protocol=snmp. On clusters without ICE leader nodes, you can also specify an SNMP community string. If you want to specify a community string, after the snmp specification, enter a forward slash (/), and the SNMP community string. For example:

  ```
  pdu_protocol="snmp/mystring"
  ```

  On a cluster with ICE leader nodes, the protocol is MODBUS. Specify pdu_protocol=modbus.

After you install the cluster manager, configure the clmgr-power service on the PDUs. For information about how to configure the clmgr-power service, see the following:

HPE Performance Cluster Manager Administration Guide

Cluster definition file example - Configuring cooling devices (HPE Apollo 9000)

Cooling devices circulate a cooling liquid through the cluster. An HPE Apollo 9000 cluster can include either or both of the following cooling devices:

- HPE Adaptive Rack Cooling System (ARCS) components. The cluster manager supports ARCS components as a preview feature in the HPE Performance Manager 1.3.1 release.
- Cooling distribution units (CDUs).

At this time in the installation, you do not have to add any information to the cluster definition file for cooling devices. If you are reinstalling a cluster and cooling device information appears in an existing cluster definition file, that is as expected. If you want to add information to the cluster definition file, add lines for each cooling device that are similar to the example.

Example 1: The following line pertains to an ARCS component with the hostname arcs0:

```
internal_name=cooldev0, mgmt_bmc_net_name=head-bmc,
mgmt_bmc_net_macs=99:99:99:99:99:99, device_type=arcs,
hostname1=arcs0
```

Example 2: The following line pertains to a CDU with the hostname `cdu0`:

```
internal_name=cooldev1, mgmt_bmc_net_name=head-bmc,
mgmt_bmc_net_macs=99:99:99:99:99:99, device_type=cdu,
hostname1=cdu1
```

The preceding examples show the following:

- All cooling devices have a unique internal name. The internal name starts with `cooldev` and ends in a number.
- You can give each cooling device a site-defined hostname. In these examples, the hostnames represent the type of cooling device being configured and are `arcs0` and `cdu1`. If you do not specify a hostname, the cluster manager uses the internal name for the hostname.
- The `mgmt_bmc_net_macs` field contains the MAC address of the cooling device.

Later procedures explain how to run the `discover` command to put the cooling devices under cluster manager control. After you run the `discover` command, the documentation explains how to run the `cm cooling` command. This command does the following:

- It adds cooling device information to the cluster definition file.
- It enables cooling device monitoring with the power and cooling infrastructure manager (PCIM).

Cluster definition file example - Specifying a specific IP address

When you run the `discover` command for a specific component, you can specify an IP address for that component on any of the networks.

For example, the following node definition shows the parameters that you can use to define network IP address specifications for node `service0`:

```
# File specific_ip.config
# Cluster definition file for compute nodes with specific IP addresses for various networks
[templates]
name=compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, redundant_mgmt_network=no,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=disk, card_type=ipmi, baud_rate=115200, bmc_username=admin, bmc_password=admin,
data1_net_interfaces="ens1f0,ens1f1", data1_net_name="tengignet", data1_net_bonding_mode=802.3ad,
data1_net_bonding_master=bond1

[nic_templates]
template=compute, network=head, bonding_master=bond0, bonding_mode=active-backup, net_ifs="eno1"
template=compute, network=head-bmc, net_ifs="bmc0"
template=compute, network=ib-0, net_ifs="ib0"
template=compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=service101, mgmt_bmc_net_macs="20:67:7c:e4:9a:10", mgmt_net_macs="00:0f:53:21:98:11",
data1_net_macs="00:03:80:aa:bb:ca,00:03:80:aa:bb:cb", mgmt_bmc_net_ip=172.24.1.1,
mgmt_net_ip=172.23.1.1, data1_net_ip=10.10.1.1, ib_0_ip=10.148.1.1, ib_1_ip=10.149.1.1,
template_name=compute
internal_name=service102, mgmt_bmc_net_macs="20:67:7c:e4:9a:21", mgmt_net_macs="00:0f:53:21:98:22",
data1_net_macs="00:03:80:aa:bb:ab,00:03:80:aa:bb:ac", mgmt_bmc_net_ip=172.24.1.2,
mgmt_net_ip=172.23.1.2, data1_net_ip=10.10.1.2, ib_0_ip=10.148.1.2, ib_1_ip=10.149.1.2,
template_name=compute
internal_name=service103, mgmt_bmc_net_macs="20:67:7c:e4:9a:32", mgmt_net_macs="00:0f:53:21:98:33",
data1_net_macs="00:03:80:aa:bb:ea,00:03:80:aa:bb:eb", mgmt_bmc_net_ip=172.24.1.3,
mgmt_net_ip=172.23.1.3, data1_net_ip=10.10.1.3, ib_0_ip=10.148.1.3, ib_1_ip=10.149.1.3,
template_name=compute
```

After installation, you can use the `cm node set --update-ip` command to change the IP address setting as needed. For more information, see the following:

Cluster definition file example - Specifying information for a compute node with an Arm (AArch64) architecture type

If any compute nodes in the cluster are of the Arm (AArch64) architecture type, specify additional information in the cluster definition file for the nodes. For these nodes, specify the following keywords:

- image=*image_name*
- kernel=*kernel_name*
- architecture=*arch*

The following file defines non-ICE compute nodes with an Arm (AArch64) architecture:

```
# File aarch64_compute.config
# Cluster definition file for AArch64 architecture compute nodes
[templates]
name=compute, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, redundant_mgmt_network=no,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=disk, card_type=iLO, baud_rate=115200, bmc_username=ADMIN, bmc_password=ADMIN,
image=sles12sp4-arm64, kernel=4.4.73-5-default, architecture=aarch64

[nic_templates]
template=compute, network=head, bonding_master=bond0, bonding_mode=active-backup, net_ifs="eno1"
template=compute, network=head-bmc, net_ifs="bmc0"
template=compute, network=ib-0, net_ifs="ib0"
template=compute, network=ib-1, net_ifs="ib1"

[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine, mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.1,
internal_name=service1, mgmt_bmc_net_macs="20:67:7c:e4:9a:12", mgmt_net_macs="00:0f:53:21:98:13",
template_name=compute
internal_name=service2, mgmt_bmc_net_macs="20:67:7c:e4:9a:23", mgmt_net_macs="00:0f:53:21:98:24",
template_name=compute
internal_name=service3, mgmt_bmc_net_macs="20:67:7c:e4:9a:34", mgmt_net_macs="00:0f:53:21:98:35",
template_name=compute
internal_name=service4, mgmt_bmc_net_macs="20:67:7c:e4:9a:45", mgmt_net_macs="00:0f:53:21:98:46",
template_name=compute
internal_name=service5, mgmt_bmc_net_macs="20:67:7c:e4:9a:56", mgmt_net_macs="00:0f:53:21:98:57",
template_name=compute
internal_name=service6, mgmt_bmc_net_macs="20:67:7c:e4:9a:67", mgmt_net_macs="00:0f:53:21:98:68",
template_name=compute
```

Running the discover command to complete the cluster configuration

The discover command adds component information to the cluster database and assigns component IP addresses and hostnames. If needed, the discover command also performs component-specific tasks such as the following:

- Assigning an image
- Powering up compute nodes, scalable unit (SU) leader nodes, or ICE leader nodes
- Configuring management switches

Use the discover command in the following situations:

- During initial configuration. Run the discover command after you run the configure-cluster command.
- After initial installation. Use the discover command to reconfigure the IP address, hostname, MAC address, bonding more, or other setting for a component.

When you run the discover command, the cluster manager does the following:

- Adds the component to the cluster manager database.
- Completes component-specific tasks. These tasks include pushing images, switch configuration, powering up, and so on.

After you install the cluster manager, you can enable the power management software on the PDUs. To change the software settings for the PDUs, use the cm node set command. For more information, see the following:

Verifying power operations and configuring power management

For information about the discover command, enter one of the following commands:

- `# discover -h`
- `# man discover`

For discover command examples, see the following:

discover command examples that use a cluster definition file

The following figure is an example that shows the installation process flow on an HPE SGI 8600 cluster system.

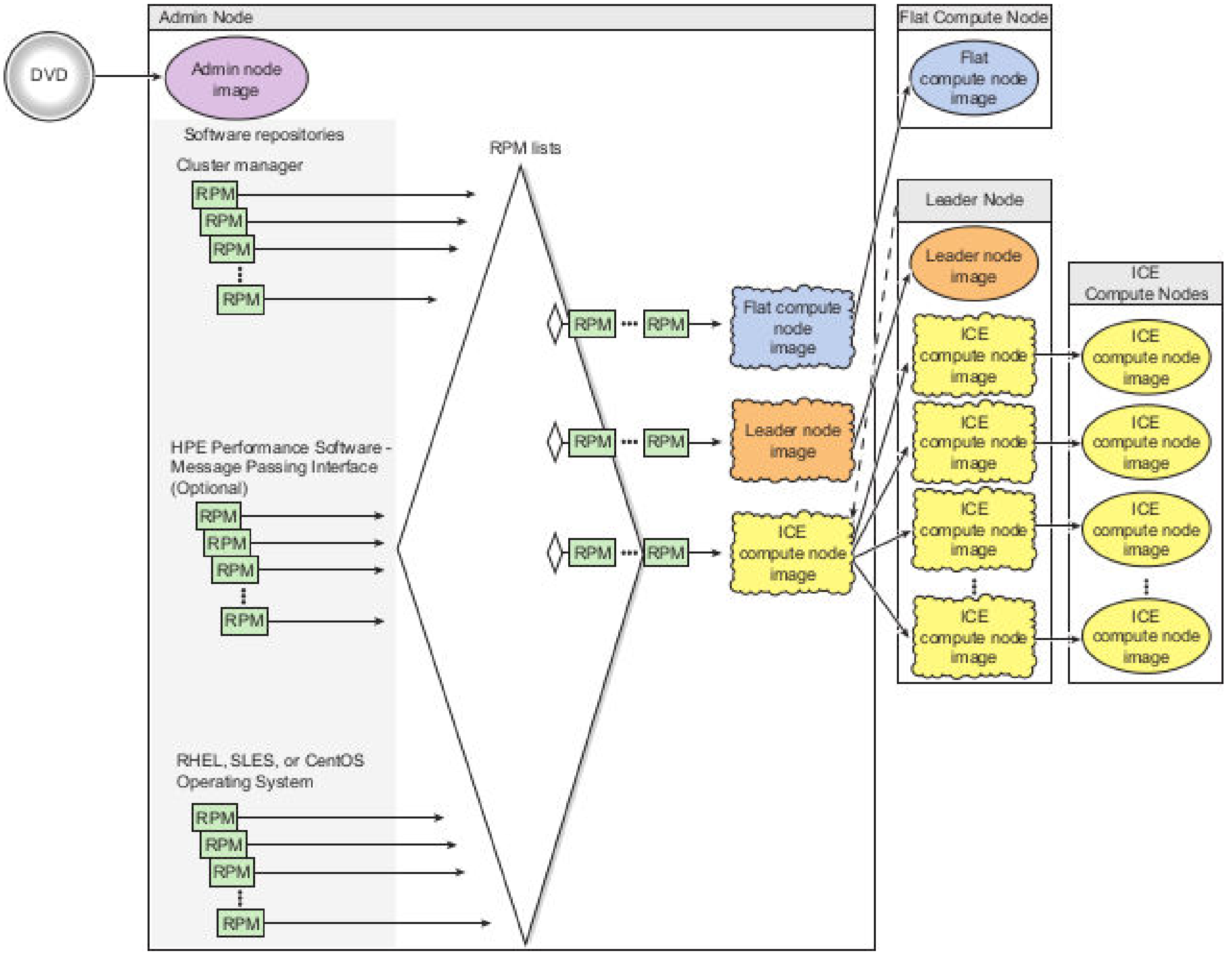


Figure 3: HPE SGI 8600 software installation process

Proceed to the following:

- To configure a cluster with scalable unit (SU) leader nodes, proceed to the following:

  Configuring a cluster with scalable unit (SU) leader nodes

- To configure a cluster without scalable unit leader nodes, proceed to the following:

  Configuring a cluster without scalable unit (SU) leader nodes

Configuring a cluster with scalable unit (SU) leader nodes

Complete the procedure in this topic if the cluster is an HPE Apollo cluster with SU leader nodes.

Procedure

1. Verify the cluster definition files.

   If you followed the procedure at the following link, you have two separate cluster definition files:

   Cluster definition file example - HPE Apollo cluster with scalable unit (SU) leader nodes

NOTE: On clusters with SU leader nodes, the management switches and SU leader nodes must be discovered, booted, and configured before you run the `discover` command to configure the compute nodes.

2. From the admin node, run the `discover` command twice.

   a. First, enter the `discover` command in the following format to update the cluster database with relevant templates:

   ```
   discover --configfile su_leader_def_file --update-templates
   ```

   b. Second, enter the `discover` command in the following format to configure all the switches and SU leader nodes defined in the cluster definition file:

   ```
   discover --configfile su_leader_def_file --all
   ```

   For example:

   ```
   # discover --configfile mgmtsw_suleader.config --update-templates
   # discover --configfile mgmtsw_suleader.config --all
   ```

3. (Optional) Use the `cm node console` command to monitor the PXE boot process on one or more SU leader nodes.

   For example:

   ```
   # cm node console -n leader1
   ```

4. Verify that all the SU leader nodes have booted.

   Use the `cm power status` command in the following format:

   ```
   cm power status -t node hostname
   ```

   For *hostname*, specify the SU leader node hostnames. If you named them in a similar way, you can use wildcard characters.

   For example, if the hostnames are `leader1`, `leader2`, and `leader3`, you can enter the following:

   ```
   # cm power status -t node "leader[1-3]"
   leader1    BOOTED
   leader2    BOOTED
   leader3    BOOTED
   ```

5. Proceed to one of the following:

   - If the cluster has a high availability admin node, proceed to the following:

     (Conditional) Completing the switch port configuration for a high availability (HA) admin node configuration

   - If the cluster does not have an HA admin node but does have liquid cooling units, proceed to the following:

     (Conditional) Configuring liquid cooling components

   - If the cluster does not have an HA admin node and does not have liquid cooling units, proceed to the following:

     (Conditional) Completing the scalable unit (SU) leader node configuration

Configuring a cluster without scalable unit (SU) leader nodes

Complete the procedure in this topic if the cluster is one of the following:

- HPE Apollo without SU leader nodes
- HPE SGI 8600
- SGI Rackable

When you completed the procedure called "**Verifying the cluster definition file**", you created at least one cluster definition file. Example names for the cluster definition file(s) are as follows:

- `mgmtsw.config`
- `compute.config`
- `ha_leader.config`
- `aarch64_compute.config`
- `generic_component.config`
- `specific_ip.config`
- `leaders.config`

For `discover` command examples, see the following:

`discover` command examples that use a cluster definition file

Procedure

1. Through an `ssh` connection, log into the admin node as the root user.
2. Verify the cluster definition file(s).

   For examples, see the following:

   Cluster definition file examples with node templates, network interface card (NIC) templates, and predictable names
3. (Conditional) Activate the NFS compute node image.

   Complete this step if the diskless compute nodes are configured in the cluster definition file with `rootfs=nfs`.

   When you complete this step, the cluster manager assumes that you want to configure the admin node as an NFS server for diskless compute nodes. In other words, you enable the admin node to act as an NFS server to the diskless compute nodes in the cluster. When each compute node boots, it can obtain a copy of the NFS root file system to use as the compute node root file system.

   Enter the following command:

   ```
   # activate-nfs-image image
   ```

   For *image*, specify the image name.
4. (Conditional) Run the `cmcdetectd` service.

   Complete this step if the cluster includes chassis management controllers (CMCs).

   Complete the following actions:

a. Enter the following command to enable the service:

```
# systemctl enable cmcdetectd
```

b. Enter the following command to start configuring the CMCs into the cluster:

```
# systemctl start cmcdetectd
```

c. Wait for the CMCs to come up.

To confirm that the CMCs are up, do one or both of the following:

Run the following command and verify that all CMCs appear in the output:

```
# cm node show -t system chassis
r1c1
r1c2
r1c3
r1c4
```

5. Use the `discover` command in the following format to update the configuration templates in the cluster definition file:

   ```
   discover --configfile cluster_definition_file --update-templates
   ```

   For *cluster_definition_file*, specify the name of the cluster definition file on the admin node.

   For example:

   ```
   # discover --configfile compute.config --update-templates
   ```

6. Use the `discover` command in the following format to configure the components named in the cluster definition file:

   ```
   discover cluster_definition_file --all
   ```

   For *cluster_definition_file*, specify the name of the cluster definition file on the admin node.

   For example:

   ```
   # discover --configfile compute.config --all
   ```

7. (Optional) Use the `cm node console` command to monitor the PXE boot process on one or more compute nodes.

   This command has the following format:

   ```
   cm node console -n hostname
   ```

   For *hostname*, specify the hostname of one of the nodes in the cluster definition file. For example, `service1`.

8. (Optional) Monitor the switch configuration process.

   If management switches or components that require management switch configuration were configured, enter the following command to monitor the progress of the switch configuration:

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

9. Verify that all nodes booted.

   Enter one or more `cm power status` commands. For example, enter the following command to verify the boot status of the service nodes:

   ```
   # cm power status -t node "service*"
   ```

10. Use the `switchconfig` command to change the management switch password for the `admin` account.

    The format for this command is as follows:

    ```
    switchconfig change_password --switches hostname --new new_password
    ```

    The variables in the command are as follows:

    - For *hostname*, specify the hostname of the management switch.
    - For *new_password*, specify a strong, new password for the switch.

    For example:

    ```
    # switchconfig change_password --switches mgmtsw0 --new Hp3@dm!n2o20
    ```

    NOTE: HPE strongly recommends that you implement standard and secure practices to store all passwords at your site. Do not lose this information.

11. Enter the following command to save the changed configuration to the nonvolatile memory (NVM) on the switches:

    ```
    # switchconfig config -s all --save
    ```

12. Proceed to one of the following:

    - If the cluster has a high availability (HA) admin node, proceed to the following:

      (Conditional) Completing the switch port configuration for a high availability (HA) admin node configuration

    - If the cluster does not have an HA admin node but does have liquid cooling units, proceed to the following:

      (Conditional) Configuring liquid cooling components

    - If the cluster does not have an HA admin node and does not have liquid cooling units, proceed to the following:

      Backing up the cluster

`discover` command examples that use a cluster definition file

The following topics show how to use the `discover` command with a cluster definition file. In these examples, the `discover` command format is always the following:

```
discover --configfile cluster_definition_file [--arg1 value] [--arg2 value] ...
```

In this format, the `discover` command reads the *cluster_definition_file* and adds one, several, or all cluster components defined in the file to the cluster database. During an initial installation, if you run the `discover` command multiple times, the required discovery order is as follows:

- Management switches.
- Scalable unit (SU) leader nodes, if present.
- All other node types. This list includes rack leader controllers (RLCs), ICE compute nodes, non-ICE compute nodes, PDUs, and other components.

`discover` command example - retrieving cluster definition file information

The following command retrieves the current cluster configuration and writes the configuration information to `stdout`:

```
discover --show-configfile [--images] [--kernel] [--bmc-info] [--ips] [--skip-examples] [--kernel-parameters]
```

If you specify any command parameters, the file includes or excludes information as follows:

Parameter	Effect on output
`--images`	Includes image information
`--kernel`	Includes kernel information
`--bmc-info`	Includes management card/BMC information such as the username, password, and baud rate information
`--ips`	Includes the IP address assigned to each component
`--skip-examples`	Suppresses example templates
`--kernel-parameters`	Includes kernel parameters for each node

`discover` command example - updating templates in the cluster database

The `[templates]` section of the cluster definition file lets you define node characteristics for a group of nodes. If you edit the `[templates]` section or the `[nic_templates]` sections, enter the `discover` command in the following format to update the cluster database:

```
discover --update-templates --configfile config_file_name
```

For *config_file_name*, specify the name of the configuration file you need to update.

For example:

```
discover --update-templates --configfile compute.config
```

`discover` command example - configuring one, several, or all components

You can use a single `discover` command to configure one component, multiple components, or all cluster components. The following examples show these methods:

- Configuring all management switches

  The following command adds all management switches named in `mgmtsw.config` to the cluster:

  ```
  # discover --configfile mgmtsw.config --all
  ```

- Configuring all management switches and scalable unit (SU) leader nodes

  The following command adds all management switches and all SU leader nodes named in `mgmtsw_suleader.config` to the cluster:

  ```
  # discover --configfile mgmtsw_suleader.config --all
  ```

- Configuring all non-ICE compute nodes under an SU leader node

  The following command adds all compute nodes that are under an SU leader node named in `su_compute.config` to the cluster:

  ```
  # discover --configfile su_compute.config --all
  ```

- Configuring one management switch

The following command adds a single management switch, named mgmtsw0, to the cluster. The switch has an entry in the cluster definition file called mgmtsw.config. The command is as follows:

```
# discover --configfile mgmtsw.config --mgmtsw 0
```

- Configuring one non-ICE compute node

  The following command adds one non-ICE compute node, named service1, to the cluster. The node has an entry in the cluster definition file called compute.config. The command is as follows:

  ```
  # discover --configfile compute.config --node 1
  ```

- Configuring multiple compute nodes

  The following command adds ten non-ICE compute nodes, named service1 through service10, to the cluster. The nodes have entries in the cluster definition file called compute.config. The command is as follows:

  ```
  # discover --configfile compute.config --nodeset 1,10
  ```

- Configuring one ICE leader node

  The following command adds one ICE leader node, named r1lead, to the cluster. The node has an entry in the cluster definition file called leaders.config. The command is as follows:

  ```
  # discover --configfile leaders.config --leader 1
  ```

- Configuring multiple ICE leader nodes

  The following command adds ten ICE leader nodes, named r1lead through r10lead, to the cluster. The nodes have entries in the cluster definition file called leaders.config. The command is as follows:

  ```
  # discover --configfile leaders.config --leaderset 1,10
  ```

- Configuring one power distribution unit (PDU)

  The following command adds one PDU, named pdu1, to the cluster. The node has an entry in the cluster definition file called pdu.config. The command is as follows:

  ```
  # discover --configfile pdu.config --pdu 1
  ```

discover command examples that do not use a cluster definition file

HPE strongly recommends that you use a cluster definition file to add nodes to the cluster management software. However, if absolutely needed, use the information in the topics that follow to add various components to the cluster without a cluster definition file.

Without a cluster definition file, you run the discover command directly from a command line session on the admin node. Because this is run from a command line session, if a parameter requires a comma-separated list, do the following:

- Put the list into quotation marks (" ")
- Escape the quotation marks with a slash (\).

The examples in the following topics contain more information. In general, the command uses the following format to string together mandatory pieces of information:

```
discover --component num,parameter1,parameter2,parameter3,...,parameterX
```

For a complete list of discover parameters, do one of the following:

- Enter the following command:

  ```
  # discover --help
  ```

- View the `discover`(8) manpage.

Assigning non-ICE compute nodes to a networking group

To enable monitoring from the cluster manager GUI, assign each non-ICE compute node to a network group. You can group the nodes into network groups in one of the following ways:

- From the GUI after the cluster manager is installed
- From the `discover` command when you add nodes to the cluster, when you initially install the cluster, or when you decide to assign nodes to network groups

A network group can include up to 288 non-ICE compute nodes. Make sure that all the nodes you configure into a network group are all attached to the same switch.

Do not attempt to add non-ICE compute nodes to the admin network group or the admin system group. The admin network group and the admin system group are reserved for the admin node only.

The following `discover` command adds non-ICE compute nodes `n1` through `n64` to the network group called `rack1nodes`:

```
# discover --network_group=rack1nodes --nodeset 1,64
```

The preceding command creates the network group called `rack1nodes` if it does not already exist.

Configuring a single component

The `discover` commands used to configure individual components are all very similar. The following commands show the beginnings of `discover` commands for the components you can configure on the command line:

- For scalable unit (SU) leader nodes and non-ICE compute nodes:

  ```
  discover --node X,...
  ```

- For ICE leader nodes:

  ```
  discover --leader X,...
  ```

- For management switches:

  ```
  discover --mgmtswitch X,...
  ```

- For power distribution units (PDUs):

  ```
  discover --pdu X,...
  ```

- For InfiniBand switches:

  ```
  discover --ibswitch X,...
  ```

Example 1. The following command adds a single non-ICE compute node to the cluster. Note the various parameters that are defined.

```
# discover --node 1,hostname1=n1,mgmt_bmc_net_macs=00:11:22:33:44:44,mgmt_net_macs=00:11:22:33:44:45,\
mgmt_net_ip=172.23.1.1,mgmt_bmc_net_ip=172.24.1.1,mgmt_net_name=head,mgmt_bmc_net_name=head-bmc,\
bmc_username=admin,bmc_password=admin,baud_rate=115200,mgmt_net_bonding_mode=active-backup,mgmt_net_interfaces=eno1,\
```

```
redundant_mgmt_network=no,rootfs=disk,conserver_logging=yes,console_device=ttyS0,dhcp_bootfile=grub2,transport=udpcast,\
switch_mgmt_network=yes
```

Example 2. The following command adds a single HPE SGI 8600 ICE leader node to the cluster. Note the various parameters that are defined. In some cases, such as in the `mgmt_net_interfaces` field, more than one value is specified. In these cases, the values are enclosed in quotation marks (`"  "`) and separated by a comma (`,`).

```
# discover --leader 1,mgmt_bmc_net_macs=00:11:22:33:aa:aa,\
\mgmt_net_macs=00:11:22:33:aa:ab/00:11:22:33:aa:ac,mgmt_net_name=head,mgmt_bmc_net_name=head-bmc,\
\bmc_username=admin,bmc_password=admin,baud_rate=115200,mgmt_net_bonding_mode=802.3ad,\
mgmt_net_interfaces=eno1/eno2,\redundant_mgmt_network=yes,rootfs=disk,conserver_logging=yes,console_device=ttyS0,\
dhcp_bootfile=grub2,transport=udpcast,\switch_mgmt_network=yes
```

Example 3. The following command adds a single Ethernet management switch assigned to a spine role to the cluster. Note the various parameters that are defined and how they differ from the previous examples.

```
# discover --mgmtswitch 0,mgmt_net_name=head,mgmt_net_macs=28:98:3a:b3:c0:bb,redundant_mgmt_network=yes,\
net=head/head-bmc,ice=yes,type=spine
```

You can use the formats in this topic for any node type. For information about configuration attributes, see the following:

Specifying configuration attributes

(Conditional) Completing the switch port configuration for a high availability (HA) admin node configuration

An HA admin node configuration includes two physical admin nodes. Both nodes need a connection to the management network. The procedure in this topic explains the following:

- Verifying that the two physical admin nodes are configured to match the management switch configuration
- Configuring switch ports for correct operation

You can assume the following cabling for the two physical admin nodes that are connected directly to `mgmtsw0`:

- On physical admin node 1:
  - `eno2` connects to `mgmtsw0` on port 1/47
  - `eno3` connects to `mgmtsw0` on port 2/47
- On physical admin node 2:
  - `eno2` connects to `mgmtsw0` on port 1/48
  - `eno3` connects to `mgmtsw0` on port 2/48

Procedure

1. Log into the first physical admin node as the root user.
2. Check the bonding mode on that physical admin node:

   ```
   # grep "Bonding Mode" /proc/net/bonding/bond0
   ```

3. Log into the second physical admin node as the root user.
4. Check the bonding mode on that physical admin node:

   ```
   # grep "Bonding Mode" /proc/net/bonding/bond0
   ```

5. Analyze the output from the `grep` commands for the two nodes.

   If the output is either of the following, the switch ports are set correctly, and you can proceed to Step **8**:

   ```
   Bonding Mode: IEEE 802.3ad Dynamic link aggregation      # Signifies LACP bonding mode

   Bonding Mode: fault-tolerance (active-backup)            # Signifies Active-Backup bonding mode
   ```

   If the output is not similar to the preceding output, proceed to Step **6**.

6. Log into the virtual admin node as the root user.
7. Use the `switchconfig` command to configure the management switches.

   Example 1. Enter the following commands on an HPE Apollo 9000 cluster with LACP/802.3ad bonding:

   ```
   # switchconfig set -s mgmtsw0 --ports 1/47 --default-vlan 1 --bonding lacp --redundant yes
   # switchconfig set -s mgmtsw0 --ports 1/48 --default-vlan 1 --bonding lacp --redundant yes
   ```

Example 2. Enter the `switchconfig` commands on an HPE SGI 8600 cluster with LACP/802.3ad bonding and include cooling VLAN 3:

```
# switchconfig set -s mgmtsw0 --ports 1/47 --default-vlan 1 --bonding lacp --redundant yes --vlans 3
# switchconfig set -s mgmtsw0 --ports 1/48 --default-vlan 1 --bonding lacp --redundant yes --vlans 3
```

NOTE: If the underlying, physical admin nodes are using `active-backup` bonding, adjust the bonding mode parameter to `--bonding none`.

8. Proceed to one of the following:

- If the cluster has scalable unit (SU) leader nodes, proceed to the following:

  (Conditional) Completing the scalable unit (SU) leader node configuration

- If the cluster does not have SU leader nodes, proceed to the following:

  Backing up the cluster

(Conditional) Configuring liquid cooling components

The following types of clusters include liquid cooling components:

- HPE Apollo 9000 clusters
- HPE SGI 8600 clusters
- SGI ICE XA clusters

Clusters can include one or more of the following cooling component types:

- HPE Adaptive Rack Cooling Systems (ARCS). Used on HPE Apollo 9000 clusters. The cluster manager supports ARCS components as a preview feature in the HPE Performance Manager 1.3.1 release.
- Cooling distribution units (CDUs). Used on HPE Apollo 9000 clusters and HPE SGI 8600 clusters.
- Cooling rack controllers (CRCs). Used on HPE SGI 8600 clusters.

If the cluster includes any of the preceding cooling components, you can use cluster manager tools to view cooling component alerts. Complete one or more of the following procedures to enable viewing of cooling component alerts:

- **Configuring an HPE Adaptive Rack Cooling System (ARCS) component on an HPE Apollo 9000 cluster**
- **Configuring a cooling distribution unit (CDU) (HPE Apollo 9000)**
- **Configuring liquid cooling components on HPE SGI 8600 clusters and SGI ICE XA clusters**

Configuring an HPE Adaptive Rack Cooling System (ARCS) component on an HPE Apollo 9000 cluster

After this procedure is complete, the power and cooling infrastructure manager (PCIM) is enabled. You can use PCIM to monitor the cooling components. For more information about PCIM, see the following:

HPE Performance Cluster Manager Administration Guide

Procedure

1. Log in as the root user to the admin node.
2. Obtain the MAC address of the ARCS component.

   If necessary, complete the procedure in the following topic, and return here when you have the MAC address:

   Using the `switchconfig` command to determine the MAC address for a cooling component
3. Enable the ARCS component.

   Use the `cm cooldev arcs add` command in one of the following formats to enable the ARCS component:

   - Format 1 - Adds the ARCs component to the cluster based on its MAC address:

     ```
     cm cooldev arcs add -m component_mac_addr -n hostname [-i ip_addr]
     ```

Use this command format the first time an ARCS component is added to the cluster. This command requires you to provide the MAC address and a hostname.

- Format 2 - Adds the ARCS component to the cluster using a previously assigned IP address:

  ```
  cm cooldev arcs add -n hostname -i ip_addr
  ```

  Use this format, if the IP address was statically configured, is reachable, and is active on the ARCS component.

The variables are as follows:

Variable	Specification
component_mac_addr	The MAC address of the component. If the command fails to configure the MAC address you specify, see the `cm cooldev cdu add` help output for information about specifying the `--Interface` *NIC* parameter.
hostname	The hostname that you want to assign to the cooling component, or the hostname that is active on the component.
ip_addr	In Format 1, you can specify an IP address, as follows: • If you specify an IP address, make sure it is an active IP address. Such an IP address might have been assigned statically. • If you do not specify an IP address, the cluster manager assigns an IP address, configures that IP address in DHCP, and enables the CDU to obtain that IP address. In Format 2, you do not specify the cooling component MAC address, so specify a statically assigned *ip_addr* address. This IP address is required to be active. In this case, it is assumed that the MAC address is already in the cluster database. You might use this format for a reinstallation or if you need to add the CDU to the cluster again after a maintenance period or outage.

For more information about the commands to add, delete, or display CDUs, see the manpages for these commands or enter one or more of the following:

```
# cm cooldev arcs -h
# cm cooldev arcs add -h
# cm cooldev arcs delete -h
# cm cooldev arcs show -h
```

4. Proceed as follows:

If you have more than one ARCS component, repeat the preceding steps for the additional component(s).

When all are configured, proceed to one of the following:

- To configure a cooling distribution unit (CDU), proceed to the following:

  Configuring a cooling distribution unit (CDU) (HPE Apollo 9000)

- If the cluster does not contain CDUs, proceed to the following:

  (Conditional) Completing the scalable unit (SU) leader node configuration

Configuring a cooling distribution unit (CDU) (HPE Apollo 9000)

After this procedure is complete, the power and cooling infrastructure manager (PCIM) is enabled. You can use PCIM to monitor the cooling components. For more information about PCIM, see the following:

HPE Performance Cluster Manager Administration Guide

For information about how to configure a CDU for an HPE SGI 8600 or SGI ICE XA cluster, see the following:

Configuring liquid cooling components on HPE SGI 8600 clusters and SGI ICE XA clusters

Procedure

1. Log in as the root user to the admin node.
2. Obtain the MAC address of the CDU.

   If necessary, complete the procedure in the following topic, and return here when you have the MAC address:

   Using the `switchconfig` command to determine the MAC address for a cooling component
3. Enable the CDU.

   Use the `cm cooldev cdu add` command in one of the following formats to enable the CDU component:

   - Format 1 - Adds the CDU to the cluster based on its MAC address:

     ```
     cm cooldev cdu add -m component_mac_addr -n hostname [-i ip_addr]
     ```

     Use this command format the first time a CDU is added to the cluster. This command requires you to provide the MAC address and a hostname.
   - Format 2 - Adds the CDU to the cluster using a previously assigned IP address:

     ```
     cm cooldev cdu add -n hostname -i ip_addr
     ```

     Use this format if the IP address was statically configured, is reachable, and is active on the CDU.

   The variables are as follows:

Variable	Specification
component_mac_addr	The MAC address of the component. If the command fails to configure the MAC address you specify, see the `cm cooldev cdu add` help output for information about specifying the `--Interface` *NIC* parameter.
hostname	The hostname that you want to assign to the cooling component, or the hostname that is active on the component.
ip_addr	In Format 1, you can specify an IP address, as follows: • If you specify an IP address, make sure it is an active IP address. Such an IP address might have been assigned statically. • If you do not specify an IP address, the cluster manager assigns an IP address, configures that IP address in DHCP, and enables the CDU to obtain that IP address. In Format 2, you do not specify the cooling component MAC address, so specify a statically assigned *ip_addr* address. This IP address is required to be active. In this case, it is assumed that the MAC address is already in the cluster database. You might use this format for a reinstallation or if you need to add the CDU to the cluster again after a maintenance period or outage.

For more information about the commands to add, delete, or display CDUs, see the manpages for these commands or enter one or more of the following:

```
# cm cooldev cdu -h
# cm cooldev cdu add -h
# cm cooldev cdu delete -h
# cm cooldev cdu show -h
```

4. Proceed as follows:

If you have more than one CDU, repeat the preceding steps for the additional CDU(s).

When all are configured, proceed to one of the following:

- If the cluster has scalable unit (SU) leader nodes, proceed to the following:

  (Conditional) Completing the scalable unit (SU) leader node configuration

- If the cluster does not have scalable unit leader nodes, proceed to the following:

  Backing up the cluster

Using the `switchconfig` command to determine the MAC address for a cooling component

Procedure

1. Log into the admin node as the root user.
2. Obtain network information for the cluster or plan to visually inspect the components and cabling.

   Proceed as follows:

   - If you have network information, such as the spreadsheet used for the cluster when it was manufactured at the factory, proceed to Step **3**.
   - If you do not have network information, you need to visually inspect the cluster. Proceed to Step **4**.

3. Examine the network information for the cluster.

   If the cluster was assembled at the factory, a network spreadsheet is available. If necessary, contact your HPE representative to obtain a copy. From the spreadsheet, determine the following:

   - The hostname of the switch into which the cooling component is plugged.
   - The switch port for the cable that attaches the cooling component to the cluster.

   Proceed to Step **7**.

4. Enter the following command to retrieve the hostnames for all the switches in the cluster:

   ```
   # cm group system show mgmt_switch
   mgmtsw0
   mgmtsw1
   mgmtsw100
   mgmtsw101
   mgmtsw102
   mgmtsw103
   mgmtsw104
   mgmtsw105
   mgmtsw2
   ```

   This command shows you how many switches are in the cluster and the hostnames of the switches. You might find this information useful when completing the rest of the steps in this procedure.

5. Check the labels on the cables going into each switch.

   Example labels are in the **Cable label** column of the following table:

Cable label	Orientation	Derived hostname
SW0A	Top switch, ports 1/0/X	`mgmtsw0`
SW0B	Bottom switch, ports 2/0/X	`mgmtsw0`

Table Continued

Cable label	Orientation	Derived hostname
SW1A	Top switch, ports 1/0/X	`mgmtsw1`
SW1B	Bottom switch, ports 2/0/X	`mgmtsw1`

As you can see, the you can derive the hostname for each switch by examining the labels on the cables.

6. Find the cable that connects the switch and the cooling unit.

   Note the port number on the switch that the cable plugs into.

7. Enter the `switchconfig` command in the following format:

   ```
   switchconfig info -s mgmtsw --fdb
   ```

   For *mgmtsw*, specify the hostname of the management switch that the cooling component is plugged into.

   For example:

   ```
   # switchconfig info -s mgmtsw1 --fdb
   ```

8. Analyze the output from the `switchconfig` command.

   In the `switchconfig` command output, find the line for the cooling component port in the switch.

   For example, assume that the cooling component is plugged into switch port 12. In the following output, the line for port 12 is highlighted. The information for the MAC address is in column 1. Properly formatted, the MAC address is `78:04:73:2f:a7:13`.

   ```
   # switchconfig info -s mgmtsw1 --fdb
   ==== L2 FDB(mac-address-table) Table Information on mgmtsw1 ====

   Running command - `display mac-address`...

           MAC Address      VLAN ID    State            Port/NickName            Aging
           2067-7ce4-f31c   1          Learned          GE1/0/7                  Y
           2067-7ce4-f336   1          Learned          GE1/0/3                  Y
           2067-7ce4-f34c   1          Learned          GE1/0/5                  Y
           48df-3787-a820   1          Learned          BAGG125                  Y
           48df-3787-d080   1          Learned          BAGG125                  Y
           48df-3789-4590   1          Learned          BAGG125                  Y
           7804-732f-a713   1          Learned          GE1/0/12                 Y
           98f2-b3ea-244f   1          Learned          BAGG111                  Y
           d4c9-efcf-b186   1          Learned          BAGG111                  Y
           ec9b-8b60-7ea6   1          Learned          BAGG125                  Y
           ec9b-8b60-7eb0   1          Learned          BAGG125                  Y
           ec9b-8b60-7ea6   1998       Learned          BAGG125                  Y
           ec9b-8b60-7ebd   1998       Learned          BAGG125                  Y
   ```

9. Return to one of the following procedures:

   To configure an HPE Adaptive Rack Cooling System (ARCS) component, proceed to the following:

   Configuring an HPE Adaptive Rack Cooling System (ARCS) component on an HPE Apollo 9000 cluster

   To configure a cooling distribution unit (CDU), proceed to the following:

   Configuring a cooling distribution unit (CDU) (HPE Apollo 9000)

Configuring liquid cooling components on HPE SGI 8600 clusters and SGI ICE XA clusters

For information about the IP addresses for cooling distribution units (CDUs) and cooling rack controllers (CRCs), see the following:

Liquid cooling cell network IP addresses (HPE SGI 8600 clusters)

The following procedure explains how to configure the switches attached to the liquid cooling components. This procedure configures the liquid cooling components into the cluster.

Procedure

1. Gather information about the liquid cooling component switches in the cluster.

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

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

3. Retrieve information about the virtual local area networks (VLANs) that are configured at this time.

   For example:

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

   The preceding output shows that the liquid cooling component VLAN is VLAN 3.

4. Use the `switchconfig set` command to configure the CDU and CRC ports that connect to the liquid cooling units.

   The following format shows the required parameters:

   ```
   switchconfig set -b none -d mcell_vlan -p ports -s switch
   ```

   Enter an individual `switchconfig set` command for each switch on the cluster network.

   The variables are as follows:

Variable	Specification
mcell_vlan	The VLAN number of the liquid cooling unit network. For *mcell_vlan*, use the output from the `cattr list` command as shown earlier in this procedure. The default is 3. Hewlett Packard Enterprise recommends that you do not change this value.
ports	The target ports. The command configures the target ports that are connected to the cooling equipment.
switch	The ID number of the management switch to which the component is attached. For example: `mgmtsw0`. To determine this value, visually inspect the switch, as follows: • Locate each CDU or CDC. The following are example labels for CDUs: `DU01`, `DU02`, and so on. • Follow the cable that connects each CDU or CDC to a switch. The following is an example label for a cable that connects each CDU to a switch: `DU01-LAN1 | 101MSW0A-36`. • Review the label on the switch. Make sure that the labels on the cables correspond to the labels on the switch ports.

For example, the following command configures VLAN 3 on management switch 0 for target ports 1:34, 1:35, and 1:36:

```
switchconfig set -b none -d 3 -p 1:34,1:35,1:36 -s mgmtsw0
```

NOTE: If you make a mistake in your configuration, you can reset the ports back to the default settings. The following example command removes the configuration of VLAN 3 from the target ports:

```
switchconfig unset -p 1:34,1:35,1:36 -s mgmtsw0
```

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

   Step **4**

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

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

   NOTE: This step is important. If a power outage or other interruption occurs, the switch stack boots with the saved configuration.

   Use the `switchconfig` command in the following format:

   ```
   switchconfig config --save -s mgmtsw0[,mgmtsw1,mgmtsw2,...]
   ```

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

7. Proceed to the following:

   Backing up the cluster

(Conditional) Completing the scalable unit (SU) leader node configuration

To complete the configuration on a cluster with SU leader nodes, complete the following procedures:

- **Creating the scalable unit (SU) leader configuration file**
- **Obtaining or creating a scalable unit (SU) leader node list file**
- **Configuring the Gluster file system and completing the cluster configuration**

Creating the scalable unit (SU) leader configuration file

Procedure

1. Specify the number of SU leader nodes in the SU leader node setup file.

    a. Open the SU leader node setup file with a text editor. This file resides at the following location:

    ```
    /opt/clmgr/etc/su-leader-setup.conf
    ```

    b. Verify that the password fields in the setup file are correct for this cluster.

    The defaults for these fields are as follows:

    ```
    password=cmdefault
    bmc_user=admin
    bmc_password=admin
    ```

    c. Save and close the SU leader node setup file.

2. Proceed to the following:

    Obtaining or creating a scalable unit (SU) leader node list file

Obtaining or creating a scalable unit (SU) leader node list file

The procedure in this topic explains how to complete the following file and write it to the correct location:

```
/opt/clmgr/etc/su-leader-nodes.lst
```

Procedure

1. Determine whether or not you have a usable copy of the SU leader node list file, `su-leader-nodes.lst`.

    On a configured system, this file resides in the following location:

    ```
    /opt/clmgr/etc/su-leader-nodes.lst
    ```

    Your situation can be one of the following:

    - You have a copy that you can use.

A copy you can use contains information for the current cluster configuration. All hardware in use at this time is reflected in the cluster configuration. This can be the original SU leader node list file or it can be a file from an on-site backup location.

If you added hardware after you took delivery of the cluster, make sure the file includes updates for that new hardware.

If you have a current, intact copy you can use, you do not need to complete the rest of this procedure. Write the file to the following location:

```
/opt/clmgr/etc/su-leader-nodes.lst
```

After you write the file to its location, proceed to the following:

Configuring the Gluster file system and completing the cluster configuration

- You do not have a copy that you can use.

  In this case, you might have a file that does not include updates for new hardware that you added after you took delivery of the cluster from the factory. In another case, you might have changed the cluster hardware after the date of your last backup copy. If you have no copy at all, you can obtain a copy of the original SU leader node list file from the HPE factory and update that file.

  In any case, if you do not have a copy that you can use, continue to the next step in this procedure.

2. Gather information for the IP addresses you need to specify in the SU leader node list file, which is `su-leader-nodes.lst`.

Your goal in this step is to choose unused, unique IP addresses for the following:

- An IP address on the BMC network for each SU leader node. This IP address enables the Gluster file system to communicate with the other SU leader nodes.
- An IP address for each SU leader node that the Gluster file system can use for failover. This IP address allows a Gluster resource to move to a different node.

a. Create a table (or use the following table) to record SU leader node hostnames and IP addresses:

SU leader node hostname	IP address on the head BMC network (mgmt-bmc)	IP address on the head network (mgmt)	Path to LUN
Example: `su-leader1`	`172.24.255.241`	`172.23.255.241`	`/dev/disk/by-path/pci-0000:08:00.0-scsi-0:2:0:0`
Example: `su-leader2`	`172.24.255.242`	`172.23.255.242`	`/dev/disk/by-path/pci-0000:08:00.0-scsi-0:2:0:0`
Example: `su-leader3`	`172.24.255.243`	`172.23.255.243`	`/dev/disk/by-path/pci-0000:08:00.0-scsi-0:2:0:0`

Table Continued

SU leader node hostname	IP address on the head BMC network (mgmt-bmc)	IP address on the head network (mgmt)	Path to LUN

b. Enter the following command from the admin node:

```
# configure-cluster
```

c. On the **Using cached values for the admin node network configuration ...** popup, click **OK**.

d. On the **Main Menu** screen, select **I Initial Setup Menu**, and click **OK**.

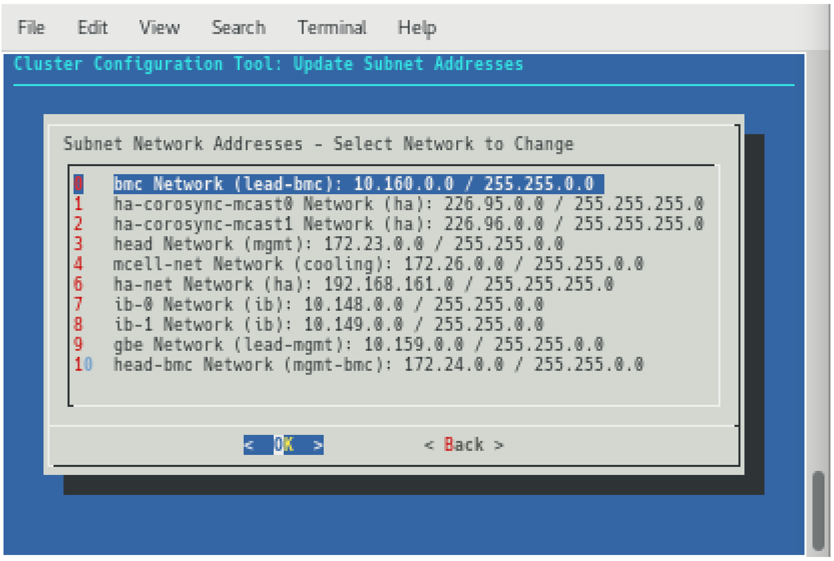
e. On the **All the steps in the following menu need to be completed in order ...** popup, click **OK**.

f. On the **Initial Cluster Setup Tasks - All Required** screen, click **N Network Settings**, and click **OK**.

g. On the **Cluster Network Settings** screen, select **S List and Adjust Subnet Addresses**, and click **OK**.

h. On the **Caution: You can adjust the subnet only for undiscovered nodes ...** popup, click **OK**.

i. On the **Subnet Network Addresses - Select Network to Change** screen, observe the addresses shown. For example:



```
Cluster Configuration Tool: Update Subnet Addresses

Subnet Network Addresses - Select Network to Change

0   bmc Network (lead-bmc): 10.160.0.0 / 255.255.0.0
1   ha-corosync-mcast0 Network (ha): 226.95.0.0 / 255.255.255.0
2   ha-corosync-mcast1 Network (ha): 226.96.0.0 / 255.255.255.0
3   head Network (mgmt): 172.23.0.0 / 255.255.0.0
4   mcell-net Network (cooling): 172.26.0.0 / 255.255.0.0
6   ha-net Network (ha): 192.168.161.0 / 255.255.255.0
7   ib-0 Network (ib): 10.148.0.0 / 255.255.0.0
8   ib-1 Network (ib): 10.149.0.0 / 255.255.0.0
9   gbe Network (lead-mgmt): 10.159.0.0 / 255.255.0.0
10  head-bmc Network (mgmt-bmc): 172.24.0.0 / 255.255.0.0
```

Figure 4: Select Network Addresses screen

j. For each SU leader node, choose an IP address on the head BMC network (mgmt-bmc). Line **10** on the screen shows the address range and the bitmask you can choose from.

Note the IP addresses you choose in the table at the beginning of this step.

k. For each SU leader node, choose an IP address on the head network (mgmt). Line **3** on the screen shows the address range and the bitmask you can choose from.

Note the IP addresses you choose in the table at the beginning of this step.

l. Click **Back** as required to return to the main menu screen, and then click **Quit**.

3. Complete the following steps to specify the SU leader node information in the list file:

a. Open the following file with a text editor:

`/opt/clmgr/etc/su-leader-nodes.lst`

b. Specify the IP addresses for each SU leader node.

Use the following to specify the SU leader node data:

- The example in the file itself
- The information you wrote in the table of step **2.a**

Use one line for each SU leader node. The line includes the following fields, separated with a comma:

`SU_leader_hostname,BMC_NIC_IP_address,management_IP_address,path_to_LUN`

For each SU leader node, specify the SU leader node hostname and then specify the information you obtained about the IP addresses in the preceding steps. You do not need to specify the *path_to_LUN* information at this time.

c. Save and close the SU leader node list file.

4. Use the `ssh` command to log in to one of the SU leader nodes.

Your goal is to select a disk to use for the shared storage that all SU leader nodes can access.

If all the SU leader nodes are of the same hardware type, you only have to analyze the storage attached to one SU leader node. Use the `ssh` command to log into that node at this time.

If the SU leader nodes are not all the same hardware type, plan to analyze the disks for each SU leader node hardware type. At this time, use the `ssh` command to log into one of the SU leader nodes.

5. Select a disk for the shared Gluster storage.

The following steps show how to choose a disk for one SU leader node.

a. Use the `lsblk` command to list the available disks on the system.

For example:

```
leader1#:~ # lsblk --paths --output NAME,MOUNTPOINT
NAME          MOUNTPOINT
/dev/sda
├─/dev/sda1
├─/dev/sda2   [SWAP]
├─/dev/sda3
├─/dev/sda11  /boot
├─/dev/sda12
├─/dev/sda21  /boot/efi
├─/dev/sda22
```

```
├─/dev/sda31 /
└─/dev/sda32 /mnt
/dev/sdb
```

Choose an empty, unmounted drive for an SU leader node. The disk drive or volume you choose cannot be associated with any other file system. The volume or drive for Gluster must be available to Gluster for its exclusive use.

The preceding output shows that `sda` is used and `sdb` is empty.

b. Enter the following command to change to the `by-path` directory:

```
# cd /dev/disk/by-path
```

c. Enter the following command to list files that end in `sdb`:

```
# ls -l | egrep "sdb$"
lrwxrwxrwx 1 root root  9 Jul  3 08:57 pci-0000:03:00.0-scsi-0:2:0:0 -> ../../sdb
```

The preceding output shows that disk `/dev/sdb` has the following persistent name:

`/dev/disk/by-path/pci-0000:08:00.0-scsi-0:2:0:0`

Write the name of the persistent disk name here: ____

A subsequent step explains how to specify that persistent name in the SU leader node list file. The installer puts a partition and a file system on the disk and configures Gluster to use the disk.

NOTE: The `/opt/clmgr/etc/su-leader-nodes.lst` file contains the path to the device file. This procedure explains how to specify LUNs in a `/dev/disk/by-path` style name. If the cluster has like-hardware, this path style specifies is a unique path that is most likely the same for every node.

d. Proceed as follows:

- If your SU leader nodes are all of the same hardware type, the nodes can use the same disk (same persistent name) for the Gluster file system. Continue to the next step in this procedure, which is as follows:

  Step **6**

- If your SU leader nodes are not all of the same hardware type, the nodes cannot use the same disk. You need to repeat the preceding selection steps for the other SU leader node hardware types. Continue to the following to select another disk for the other node hardware type(s):

  Step **5.a**

- If your SU leader nodes are of different hardware types, but you have selected a disk for each SU leader node hardware type, continue to the following:

  Step **6**

6. Complete the following steps to specify the SU leader node information in the list file:

a. Open the following file with a text editor:

`/opt/clmgr/etc/su-leader-nodes.lst`

b. Use the example in the file as a guide, and specify the path to the disk(s) you selected in this procedure.

c. Save and close the list file.

7. Proceed to the following:

Configuring the Gluster file system and completing the cluster configuration

The following procedure configures the Gluster file system on the scalable unit (SU) leader node shared disk and completes the cluster configuration.

Procedure

1. (Conditional) Back up data from the slot you want to install into.

   Complete this step if you have a previous configuration in the slot and you want to retain the data. This procedure removes all data from the disk. Use the backup program in use at your site.

2. From the admin node, enter the following command to run the SU leader node configuration scripts:

   ```
   # su-leader-setup [--destroy-gluster]
   ```

   This command creates partition tables, sets up high availability, configures the Gluster file system, and completes several other configuration tasks.

   If SU leader nodes were configured previously in the slot you are on, parts of the configuration scripts do not run without being forced. In this case, specify `--destroy-gluster` to clear the disk. When specified, the command completely deletes all content on the listed disk device for every node as it configures the partitions.

   Enter the following command for help:

   ```
   # su-leader-setup --help
   ```

3. Verify the configuration script results.

   a. Verify that there are Gluster volumes for all the SU leader nodes. Enter the command in this step from the admin node.

      For each SU leader node that you have, enter the following command:

      ```
      ssh SU_lead_hostname gluster volume status cm_shared
      ```

      For *SU_lead_hostname*, specify the hostname of one of the SU leader nodes. It does not matter which node *hostname* you specify on this command line. You need to run these commands on all the SU leader nodes.

      For example, enter the following command for an SU leader node named **leader1**:

      ```
      # ssh leader1 gluster volume status cm_shared
      Status of volume: cm_shared
      Gluster process                        TCP Port RDMA Port Online Pid
      ----------------------------------------------------------------------
      Brick 172.23.0.2:/data/brick_cm_shared 49152    0         Y      10289
      Brick 172.23.0.3:/data/brick_cm_shared 49152    0         Y      30961
      Brick 172.23.0.4:/data/brick_cm_shared 49152    0         Y      17849
      NFS Server on localhost                2049     0         Y      10722
      Self-heal Daemon on localhost          N/A      N/A       Y      10354
      NFS Server on 172.23.0.3               2049     0         Y      31340
      Self-heal Daemon on 172.23.0.3         N/A      N/A       Y      31026
      NFS Server on 172.23.0.4               2049     0         Y      18230
      Self-heal Daemon on 172.23.0.4         N/A      N/A       Y      17913
      ```

```
Task Status of Volume cm_shared
----------------------------------------------------------------------
There are no active volume tasks
```

In the preceding output, notice the following:

- Each brick is listed properly for each node.
- Each brick has a TCP port, is online, and has a PID.

b. For each SU leader node that you have, enter the following command:

```
ssh SU_lead_hostname ctdb status
```

For example:

```
# ssh leader1 ctdb status
Number of nodes:3
pnn:0 172.23.0.2       OK (THIS NODE)
pnn:1 172.23.0.3       OK
pnn:2 172.23.0.4       OK
Generation:370917201
Size:3
hash:0 lmaster:0
hash:1 lmaster:1
hash:2 lmaster:2
Recovery mode:NORMAL (0)
Recovery master:1
```

If the CTDB status shows anything other than `OK` or if the Gluster volumes do not look to be in the right state, run the following command from the admin node:

```
# su-leader-setup --bring-online
```

The **`--bring-online`** parameter ensures that the volumes and CTDB status are correct. When this command finishes, run the verification steps again on that node.

NOTE: You can run the `su-leader-setup --bring-online` command at any time the volume or CTDB status is not correct. For example, you could run this command after an SU leader node goes down if the node does not register itself properly with CTDB after coming back up.

c. Select one of the SU leader nodes, and enter the following command to check the assignment across all of the SU leader nodes:

```
ssh SU_lead_hostname ctdb ip
```

For *SU_lead_hostname*, specify the hostname of one of the SU leader nodes. Run this command once for each of the SU leader nodes.

For example:

```
# ssh leader1 ctdb ip
Public IPs on node 0
172.23.255.241 0     # This is leader1
172.23.255.242 1     # This is leader2
172.23.255.243 2     # This is leader3
```

The preceding example output shows the IP address aliases for each SU leader node. These are the IP addresses that the cluster manager assigned to each SU leader. In a failover, these addresses move from the failing nodes. You can use these IP addresses to log into a specific node.

d. For each SU leader node that you have, enter an `ip addr show` command to verify that there are two IP addresses for each SU leader node.

For example:

```
# ip addr show bond0 label bond0 | grep global
inet 172.23.1.1/16 brd 172.23.255.255 scope global noprefixroute bond0
inet 172.23.255.241/16 brd 172.23.255.255 scope global secondary bond0
```

4. Configure the admin node to work with the new SU leader nodes.

This step performs the following tasks:

- Ensures that required paths that the admin node uses are from shared storage
- Places mounts and bind-mounts in the `fstab` file
- Synchronizes all images to shared storage

Enter the following command:

```
# enable-su-leader
```

5. Activate the NFS image.

The `activate-nfs-image` command activates the specified image and enables the image for use by the NFS clients. The command copies the image and the required network boot files into NFS-exported locations on the shared storage.

Enter the `activate-nfs-image` command in the following format:

```
activate-nfs-image image_name
```

For *image_name*, specify the image name.

For example:

```
# activate-nfs-image rhel8.1
```

6. Use the information in one of the following hardware-specific tables to configure the non-ICE compute nodes into the cluster.

The following table applies to HPE Apollo 9000 clusters:

Step	Task
1	Enter the following command to enable the `cmcdetectd` service: `# systemctl enable cmcdetectd`
2	Enter the following command to configure the chassis management controllers (CMCs) into the cluster: `# systemctl start cmcdetectd`

Table Continued

Step	Task

3	Wait for the CMCs to come up.
	To confirm that the CMCs are up, do one or both of the following:
	Run the following command and verify that all CMCs appear in the output:
	`# cm node show -t system chassis`
	`r1c1`
	`r1c2`
	`r1c3`
	`r1c4`

4	Enter the following commands to start the `cmcinventory` service: **a.** `# systemctl enable cmcinventory` **b.** `# systemctl start cmcinventory`

The following table applies to HPE Apollo clusters that are not HPE Apollo 9000 clusters:

Step	Task

1.	Run the `discover` command.
	The format is as follows:
	`discover --configfile compute_nodes_def_file --all`
	For *compute_nodes_def_file*, specify the name of the file that defines the non-ICE compute nodes for this cluster.
	For example:
	`# discover --configfile su_compute.config --all`

7. Back up the cluster configuration.

   At this time, continue to the following procedure to back up the cluster configuration files:

   Backing up the cluster

Backing up the cluster

Complete the following procedures now and whenever you significantly change any hardware or software in the cluster:

- **Backing up the admin node**
- **Backing up the cluster configuration files**

Backing up the admin node

Use a backup program at your site to back up the admin node. Completing this procedure now, before you put your cluster into production. In this way, you ensure that you have a copy of the admin node that you can use in case a disaster occurs. Make sure to write the backup copies to a safe location on a computer that resides outside the cluster. An admin node backup protects the following:

- The cluster database
- The cluster definition file
- The node images
- The VCS source control system

NOTE: Make sure to back up the admin node regularly. Backing up the admin node protects your cluster configuration if a disk failure or other disaster occurs.

The following procedure explains how to back up the admin node.

Procedure

1. Use your site practices and site backup program to back up the cluster admin node.

   To restore the admin node, use the restore procedure for your site backup program.
2. (Optional) Use your site file restore practices to test a restore of the admin node.

Backing up the cluster configuration files

Complete this procedure now and at any other time you significantly modify the cluster. For example, repeat this procedure in the following situations:

- Changing cluster attributes.
- Adding nodes to the cluster.
- Deleting nodes from the cluster.
- Changing the software image on a node.
- Changing the kernel on a node.
- Changing the hostname of a node.
- Changing the IP address of a node.

If you have more than one slot, remember that backing up a slot by cloning the slot is not equal to backing up the admin node. Disk failures can occur.

The following procedure explains how to back up the cluster definition file and the cluster database.

Procedure

1. Enter the following command to back up the cluster definition file:

   ```
   # discover --show-configfile --images --kernel --bmc-info \
   --kernel-parameters --ips > filename
   ```

   For *filename*, specify a file name.

   For example:

   ```
   # discover --show-configfile --images --kernel --bmc-info \
   --kernel-parameters --ips > my.config.file
   ```

   NOTE: Hewlett Packard Enterprise recommends that you keep a copy of the cluster definition file on another computer system at your site.

   You need the cluster definition file in case you have to reconfigure one or more nodes. This file can be useful when troubleshooting. You also need the cluster definition file for disaster recovery. You can supply the cluster definition file as input to both the `discover` command and the `configure-cluster` command. The cluster definition file supplies the information that you would typically define by using the menus in the cluster configuration tool. When you specify a cluster configuration file as input to these commands, the commands read in the options from the file and implement them in the cluster.

   Save a new copy of the cluster definition file anytime you modify your cluster. Without a cluster definition file, to reconfigure any aspect of your cluster, you have to power on and power off each component during the configuration process. To restore the cluster definition file, copy the file from its backup location to the admin node.

2. Copy the cluster definition file to another server at your site.

3. (Conditional) Back up the custom partitioning file.

   Complete this step if you configured custom partitioning.

   The custom partitioning file resides in the following location:

   `/opt/clmgr/image/scripts/pre-install/custom_partitions.cfg`

4. (Conditional) Back up the scalable unit (SU) leader node configuration files.

   Complete this step if the cluster includes SU leader nodes.

   Back up the following files:

   - `/opt/clmgr/etc/su-leader-setup.conf`
   - `/opt/clmgr/etc/su-leader-nodes.lst`

5. 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. The cluster manager writes the backup file to the current working directory.

   For example:

   ```
   # sqlite3 /opt/clmgr/database/db/cmu.sqlite3 ".backup cmu.backup.sqlite3"
   ```

6. Enter the following command to start the cluster manager:

   ```
   # systemctl start cmu
   ```

   NOTE: In the future, to restore the cluster database and start the cluster manager, enter the following commands:

   ```
   # systemctl stop cmu
   # cp -i 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.

   ```
   # systemctl stop cmu
   # cp -i cmu.backup.sqlite3 /opt/clmgr/database/db/cmu.sqlite3
   cp: overwrite '/opt/clmgr/database/db/cmu.sqlite3'? y
   # systemctl start cmu
   ```

7. Copy the database backup file to another server at your site.

   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.

8. Enter the following command to save the changed configuration to the nonvolatile memory (NVM) on the switches:

   ```
   # switchconfig config -s all --save
   ```

9. Enter the following command to back up all the switch configuration information:

   ```
   # switchconfig config -s all --pull
   configuration file 'startup-config' copied from mgmtsw0 to 172.23.0.1
   at /opt/clmgr/tftpboot/mgmtsw_config_files/mgmtsw0/startup-config
   configuration file 'startup-config' copied from mgmtsw1 to 172.23.0.1
   at /opt/clmgr/tftpboot/mgmtsw_config_files/mgmtsw1/startup-config
   configuration file 'startup.cfg' copied from mgmtsw2 to 172.23.0.1
   at /opt/clmgr/tftpboot/mgmtsw_config_files/mgmtsw2/startup.cfg
   configuration file 'primary.cfg' copied from mgmtsw3 to 172.23.0.1
   at /opt/clmgr/tftpboot/mgmtsw_config_files/mgmtsw3/primary.cfg
   ```

   Observe the message that this command issues upon completion. This message contains the location of the backup files. By default, the message points to the files in the following directory on the admin node:

   ```
   /opt/clmgr/tftpboot/mgmtsw_config_files
   ```

10. Note the file name or names from the preceding command output. Copy each backup file from the admin node to a safe storage space at your site.

Configuring additional features

The cluster manager includes features that you might have to configure depending on your components. Additionally, there are features that are not required but might be of use on your system. For example:

- If you have a highly available admin node, make sure to complete the following procedure:

  Naming the storage controllers for clusters with a highly available (HA) admin node

- If you want to configure power management, complete the following procedure:

  Verifying power operations and configuring power management

NOTE: If you add or change anything on your cluster, remember to back up the cluster again. Use the procedures in the following:

Backing up the cluster

Configuring the GUI on a client system

The following procedure explains how to configure the GUI on a client computer outside of the cluster system. For example, you can install the client software on a laptop computer.

Procedure

1. On the client computer, verify that Java 8+ is installed.
2. Open a browser, and enter one of the following addresses for the admin node:
   - The IP address
   - The fully qualified domain name (FQDN)
3. Follow the instructions on the cluster manager splash page to download and install the GUI client.

Starting the cluster manager web server on a non-default port

Procedure

1. On the admin node, use a text editor to adjust the settings in the following file:

   ```
   /opt/clmgr/etc/cmuserver.conf
   ```
2. Open the corresponding ports in the firewall.

Customizing leader nodes and non-ICE compute nodes

You can use post-installation scripts to customize operations on non-ICE compute nodes and ICE leader nodes. The scripts can enable additional software, append data to configuration files, configure supplemental network interfaces, and perform other operations. For information about these scripts, see the following file:

```
/opt/clmgr/image/scripts/post-install/README
```

Configuring network groups for monitoring

If you want to use native monitoring, configure the non-ICE compute nodes into network groups. It is most common to configure all the non-ICE compute nodes under a common switch into a network group.

For information about how to configure network groups, see the following:

HPE Performance Cluster Manager Administration Guide

Naming the storage controllers for clusters with a highly available (HA) admin node

Complete the procedure in this topic if the cluster has an HA admin node.

The following procedure configures names for the storage controllers. The names enable you to manage them from the admin node.

Procedure

1. From the admin node, enter the following commands:

   ```
   # discover \
   --node 100,mgmt_net_macs=00:50:B0:AB:F6:EE,hostname1=unita,generic
   # discover \
   --node 101,mgmt_net_macs=00:50:B0:AB:F6:EF,hostname1=unitb,generic
   ```

   The commands in this step accomplish the following:

   - The commands configure hostnames and IP addresses for the storage controllers. These host names are `unita` and `unitb`.
   - The commands configure DHCP so that the storage devices automatically receive an IP address.

Verifying power operations and configuring power management

The power management service provides the following features:

Feature	Platforms
Power monitoring.	All cluster types with power measurement hardware.
Rack level and system level power and energy measurement.	All cluster types with rack-level power distribution unit (PDU) monitors.
Power limiting. You can limit power for the entire cluster, for specific racks or rack sets, or for individual nodes within the cluster.	HPE Apollo clusters. All nodes are required to have an iLO Advanced license. HPE SGI 8600

There are no power limiting defaults. If you set a power limit, make sure that the limit is set lower than the actual power that the node can generate. If the power limit is set higher than the amount of power that a node can generate, then the limit is not effective.

For information about power operations, see the following:

HPE Performance Cluster Manager Power Management Guide

For information about power monitoring, see the following:

HPE Performance Cluster Manager Administration Guide

Adjusting the domain name service (DNS) search order

A DNS search path lists the order of subdomains to try when you (or a program) need to translate a hostname into an IP address.

If you use DNS as the method to convert hostnames into IP addresses, you can configure the following:

- A specific subdomain is the first IP address to be resolved. In addition, you can specify more than one subdomain and the order in which each subdomain is to be searched.
- A DNS resolution specification that applies to the cluster globally or only for a specific node.

The following are examples of subdomains that you can specify:

- InfiniBand fabric IP addresses. For example, `ib0.clusterdomain`.
- Management fabric IP addresses. For example, `gbe.clusterdomain`.
- Public or external IP addresses. For example, `mycompany.com`.

The cluster manager sets the DNS search order after you run the cluster configuration tool. However, you can change the domain search order at any time after the cluster is installed and configured.

For more information, see the `resolv.conf` manpage.

The following topics include information about how to analyze, view, or configure search order:

- **Analyzing your environment**
- **Configuring the global or per-node DNS search order**
- **Retrieving the global or per-node DNS search order**

Analyzing your environment

Sometimes a host includes multiple network interfaces.

For example, on an HPE SGI 8600, a non-ICE compute node can have one interface for the management network and one interface for the data network, as follows:

```
r1i0n0.gbe         IN    A    10.159.0.4
r1i0n0.ib0         IN    A    10.148.0.9
```

In this example, Hewlett Packard Enterprise recommends that you set a DNS search order for this node to specify the preferred network. In general, the `ib` network is preferred to the management network.

A command that does not specify the subdomain of `.gbe` or `.ib0` uses the DNS search path to determine the IP address to return, as follows:

- The host lookup command returns the `ib0` IP address when the DNS search path is one of the following:
  - `ib0.clusterdomain clusterdomain`

or

  - `ib0.clusterdomain gbe.clusterdomain clusterdomain`

- The host lookup command returns the `gbe` IP address when the search path is one of the following:
  - `gbe.clusterdomain clusterdomain`

    or

  - `gbe.clusterdomain ib0.clusterdomain clusterdomain`

- If neither `ib0` nor `gbe` are in the DNS search path, the host lookup command returns the first entry in the DNS configuration file.

  When searching, specify the subdomains in the same search order as the domains are defined.

The DNS search order is more important when nodes with different interfaces try to reach each other. For example, if the admin node does not have an `ib0` interface, `gbe` needs to be first in the DNS search path for the admin node itself.

If IP address information for a node is in the `hosts` file, the system ignores the DNS search path.

For example, on an HPE SGI 8600 cluster with ICE leader nodes:

- If leader node `r1lead` has an entry of `10.159.0.4` for ICE compute node `r1i0n0`, then the leader node uses the management network to reach node `r1i0n0`. In the `hosts` file of ICE compute node `r1i0n0`, there is an entry that includes the IP address of leader node `r1lead` for the same reason.
- The cluster admin node has a `hosts` file that contains entries for the non-ICE compute nodes that it manages directly. Because the leader nodes manage ICE compute nodes, the admin node uses DNS to reach the ICE compute nodes.

The following topics explain how to view or configure the global or per-node search order:

- **Configuring the global or per-node DNS search order**
- **Retrieving the global or per-node DNS search order**

Configuring the global or per-node DNS search order

Procedure

1. Log into the admin node as the root user.
2. Use the following `cm node set` command to set the DNS resolution order:

   ```
   cm node set [-g] [-node node] --domain-search-path new_domain_search_path
   ```

   The parameters are as follows:

   - The `-g` parameter, if specified indicates that the command is global. If you specify `-g`, do not specify the `-n` *`node`* parameter.
   - For *node*, specify the hostname of the node. Specify this parameter if you want to specify a search path for only one node. Do not specify this parameter if you want to specify a global search path.
   - For *new_domain_search_path*, specify one or more domains to search. If you specify more than one domain, the cluster manager searches the domains in the order specified. Use a comma (`,`) character to separate domains.

Example 1. The following command sets a global domain search path:

```
admin:~ # cm node set -g --domain-search-path ib0.cm.americas.hpe.com,head.cm.americas.hpe.com
```

Example 2. The following command sets the domain search path for `r1lead`:

```
admin:~ # cm node set -n r1lead --domain-search-path head.cm.americas.hpe.com,ib0.cm.americas.hpe.com
```

Retrieving the global or per-node DNS search order

Procedure

1. Log into the admin node as the root user.
2. Use the following `cadmin` command to show the DNS search order:

   ```
   cadmin --show-domain-search-path [--node node]
   ```

   For *node*, specify the hostname of the node. Specify this parameter when you want to retrieve the search path for a specific node. Do not specify this parameter if you want to retrieve the global domain search path.

   Example 1. The following command retrieves the global domain search path:

   ```
   admin:~ # cadmin --show-domain-search-path
   ib0.cm.americas.hpe.com,head.cm.americas.hpe.com
   ```

   Example 2. The following command retrieves the domain search path for one node, `r1lead`:

   ```
   admin:~ # cadmin --show-domain-search-path --node r1lead
   head.cm.americas.hpe.com,ib0.cm.americas.hpe.com
   ```

Configuring a backup domain name service (DNS) server

Typically, the DNS on the admin node provides name services for the cluster. If you configure a backup DNS, the cluster can use a non-ICE compute node as a secondary DNS server when the admin node is unavailable. You can configure a backup DNS only after you run the `discover` command to configure the cluster. This feature is optional.

The following procedure explains how to configure a non-ICE compute node to act as a DNS.

Procedure

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

   ```
   # cnodes --compute
   ```

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

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

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

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

Initializing the Intel Omni-Path fabric manager (HPE SGI 8600 clusters)

After cluster installation, set up the Omni-Path fabric. On clusters with leader nodes, complete the procedure in this topic from each leader node. On clusters without leader nodes, complete the procedure in this topic from the admin node.

For information about managing Omni-Path fabric, see the following:

HPE Performance Cluster Manager Administration Guide

The following procedure explains how to initialize the Omni-Path fabric.

Procedure

1. Log into one of the following nodes as the root user:

   - On clusters with leader nodes, log into one of the leader nodes. For example, on an HPE SGI 8600 cluster, log into `r1lead`.
   - On clusters without leader nodes, log into the admin node.

2. Edit the `opafm.xml` file to enable logging.

   Complete the following steps:

   - Open the following file within a text editor:

     ```
     /etc/opa-fm/opafm.xml
     ```

   - Within `opafm.xml`, search for the following line:

     ```
     <!-- <LogFile>/var/log/fm0_log/<LogFile> --> <!-- log for this instance -->
     ```

   - Remove the following character strings:
     - `<!--`
     - `--!>`

     The deletions leave the line as follows:

     ```
     <LogFile>/var/log/fm0_log/<LogFile> <!-- log for this instance -->
     ```

     The characters you remove are XML comment characters, and they are not needed.

     Keep the file open so you can add information.

3. In another window, use the `ibstat` command to display the port globally unique identifiers (GUIDs) for the node:

   For example:

   ```
   r1lead:~ # ibstat -p
   0x001175010163ce82
   0x001175010163cea3
   ```

   The first line is the port GUID for `fm0`. The second line is the port GUID for `fm1`.

4. Add port GUIDs to the fabric manager instances in the `opafm.xml` file, as follows:

- Within `opafm.xml`, search for the following:

  `<Name>fm0</Name>`

- Within the block of text that describes `fm0`, locate the following string:

  `<PortGUID>0x0000000000000000</PortGUID>`

- Replace `0x0000000000000000` with the GUID string for `fm0`, which is the first line of `ibstat` command output.

- Within `opafm.xml`, search for the following:

  `<Name>fm1</Name>`

- Within the block of text that describes `fm1`, locate the following string:

  `<PortGUID>0x0000000000000000</PortGUID>`

- Replace `0x0000000000000000` with the GUID string for `fm1`, which is the second line of `ibstat` command output.

- Remain within the block of text that describes `fm1`, and make the following additional changes:

  - Change `<Start>0</Start>` to `<Start>1</Start>`
  - Change `<Hfi>1</Hfi>` to `<Hfi>2</Hfi>`
  - Change `<Port>2</Port>` to `<Port>1</Port>`
  - Change `<SubnetPrefix>0xfe80000000001001</SubnetPrefix>` to `<SubnetPrefix>0xfe80000000000001</SubnetPrefix>`.

NOTE: It is possible that some of the file already contains some of the correct values. In this case, verify that the correct values are present.

5. Configure the fabric topology.

   Choose the appropriate routing engine for the fabric topology.

   The default for the `opafm.xml` file is as follows:

   ```
   <RoutingAlgorithm>shortestpath</RoutingAlgorithm>
   ```

   For a hypercube or enhanced hypercube topology, use the following line:

   ```
   <RoutingAlgorithm>hypercube</RoutingAlgorithm>
   ```

   For a Fat Tree topology, use the following line:

   ```
   <RoutingAlgorithm>fattree</RoutingAlgorithm>
   ```

6. Save and close file `/etc/opa-fm/opafm.xml`.

7. Enter the following command to start the fabric manager:

   ```
   # systemctl start opafm
   ```

Setting a static IP address for the baseboard management controller (BMC) in the admin node

Complete the procedure in this topic if one or both of the following are true:

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

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

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

Method 1 -- To change from the BIOS

Use the BIOS documentation for the admin node.

Method 2 -- To change the IP address from the admin node

Procedure

1. Log into the admin node as the root user.
2. Enter the following command to retrieve the current network settings:

   ```
   # ipmitool lan print 1
   ```

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

   For example:

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

   Note the IP address in this step and decide whether this IP address is acceptable. The rest of this procedure explains how to keep this IP address or to set a different static IP address.

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

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

   The command in this step has the following effect:

   - The command specifies that the IP address on the BMC is a static IP address.
   - The command sets the IP address to the IP address that is currently assigned to the BMC.

   To set the IP address to a different IP address, proceed to the following step. If the current IP address is acceptable, you do not need to perform the next step.

5. (Conditional) Reset the static IP address.

Complete this step to set the static IP address differently from the current IP address. Enter `ipmitool` commands in the following format:

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

The variables are as follows:

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

For example, to set the IP address to 100.100.100.100, enter the following commands:

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

6. (Conditional) Repeat the preceding steps on the second admin node.

Complete this procedure again only if you want to configure a second admin node for a two-node high availability cluster.

Configuring the InfiniBand subnets

Perform the procedure in this topic as follows:

- If you have a cluster with leader nodes, you can configure the InfiniBand subnet either on a leader node or on a non-ICE compute node.

  Each cluster has two InfiniBand fabric network cards, `ib0` and `ib1`. Each subnet has a subnet manager. The subnet manager runs on a leader node or on a non-ICE compute node.
- If you have a cluster without leader nodes, you can configure the InfiniBand subnet on one of the non-ICE compute nodes. Some InfiniBand switches are configured for an InfiniBand subnet. Perform the procedure in this topic in the following situations:
  - Your switch is not preconfigured for InfiniBand.
  - You prefer to configure this service on a non-ICE compute node.

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

http://www.openfabrics.org

For more information about the InfiniBand fabric implementation, see the following:

HPE Performance Cluster Manager Administration Guide

The following procedure explains how to configure the master and the standby components and how to verify the configuration.

Procedure

1. Through an `ssh` connection, log into the admin node as the root user.
2. (Conditional) Configure the subnet manager to be compatible with the Arm (AArch64) architecture.

   Complete the following steps if the cluster is an HPE Apollo cluster with the Arm (AArch64) architecture and the Mellanox InfiniBand HCA model MCX545M-ECAN card:

   - On the admin node, open the following file in a text editor:

     ```
     /opt/sgi/var/sgifmcli/opensm-ib0.conf.templ
     ```

   - Add the following line at the end of the file:

     ```
     virt_enabled=2
     ```

   - Save and close the file.

3. Enter the following command to disable InfiniBand switch monitoring:

   ```
   # cattr set disableIbSwitchMonitoring true
   ```

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

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

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

     ```
     # configure-cluster
     ```

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

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

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

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

5. Select **A Configure InfiniBand ib0**, and select **Select**.
6. On the **Configure InfiniBand** screen, select **A Configure Topology**, and select **Select**.
7. On the **Topology** screen, select the topology your system uses, and select **Select**.

   The menu selections are as follows:

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

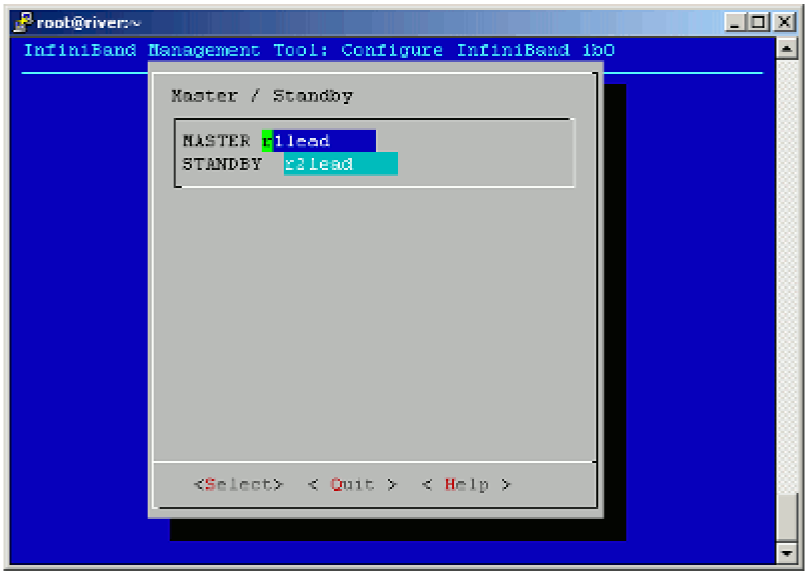
8. On the **Configure InfiniBand** screen, select **B Master / Standby**, and select **Select**.

9. On the **Master / Standby** screen, enter the component identifiers for the master (primary) and the standby (backup, secondary) subnet, and select **Select**.

   The nodes you select for the **MASTER** and **STANDBY** must have InfiniBand cards.

   Example 1. On a cluster with leader nodes, if you have only one leader node, type the leader node hostname in the **MASTER** field, and leave the **STANDBY** field blank. If you have more than one leader node, specify different leader nodes in the **MASTER** and **STANDBY** fields.

   Figure 5: Completed InfiniBand (`ib0`) Master / Standby Screen shows a completed screen for a cluster with ICE leader nodes.



```
root@river:~
InfiniBand Management Tool: Configure InfiniBand ib0

Master / Standby

MASTER r1lead
STANDBY  r2lead
```

Figure 5: Completed InfiniBand (`ib0`) Master / Standby Screen

   Example 2. On a cluster without leader nodes, enter `n1` in the **MASTER** field, and type `n101` in the **STANDBY** field.

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

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

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

11. On the InfiniBand Management Tool main menu screen, select **B Configure InfiniBand ib1**, and select **Select**.

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

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

    Select the topology that exists on your system. The menu selections are as follows:

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

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

15. On the **Master / Standby** screen, enter the component identifiers for the master (primary) and the standby (backup, secondary) subnet, and select **Select**.

Example 1. On a cluster with leader nodes, if you have only one leader node, type the leader node hostname in the **MASTER** field, and leave the **STANDBY** field blank. If you have two leader nodes, you can flip the specifications for `ib1`. For example, assume that for `ib0`, you specified **MASTER** as `r1lead` and **STANDBY** as `r2lead`. For `ib1`, you can specify **MASTER** as `r2lead` and **STANDBY** as `r1lead`. If you have three or more leader nodes, specify different leader nodes in the **MASTER** and **STANDBY** fields.

Example 2. On a cluster without leader nodes, type `n101` in the **MASTER** field, and type `n1` in the **STANDBY** field.

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

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

17. On the InfiniBand Management Tool main menu screen, select **C Administer Infiniband ib0**, and select **Select**.
18. On the **Administer InfiniBand** screen, select **Start**, and select **Select**.
19. On the **Start SM master_ib0 on ib0 succeeded!** screen, select **OK**.
20. Select **Quit** to return to the InfiniBand Management Tool main menu screen.

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

21. On the InfiniBand Management Tool main menu screen, select **D Administer Infiniband ib1**, and select **Select**.
22. On the **Administer InfiniBand** screen, select **Start**, and select **Select**.
23. On the **Start SM master_ib1 on ib1 succeeded!** screen, select **OK**.
24. On the **Administer InfiniBand** screen, select **Status**, and select **Select**.

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

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

25. Wait for the status messages to stop, and press **Enter**.
26. Select **Quit** on the menus that follow to exit the configuration tool.
27. Use the `ssh` command to change to one of the leader nodes.

    For example:

    ```
    # ssh r1lead
    ```

28. Use the `ibstatus` command to retrieve the status information for this node.

    For example, the following commands show that the node from Step **27** is connected to the fabric:

    ```
    # ibstatus
    Infiniband device 'mlx4_0' port 1 status:
        default gid:   fec0:0000:0000:0000:0002:c903:00f3:5311
        base lid:      0x1
        sm lid:        0x1
        state:         4: ACTIVE
        phys state:    5: LinkUp
        rate:          56 Gb/sec (4X FDR)
        link_layer:    InfiniBand
    ```

```
Infiniband device 'mlx4_0' port 2 status:
    default gid:  fec0:0000:0000:0001:0002:c903:00f3:5312
    base lid:     0x2a
    sm lid:       0x2a
    state:        4: ACTIVE
    phys state:   5: LinkUp
    rate:         56 Gb/sec (4X FDR)
    link_layer:   InfiniBand
```

The output shows the status as `ACTIVE` on both ports, which is correct.

29. Log into one of the nodes that is linked to the fabric, and use the `ibhosts` command to display the nodes on the fabric.

Use this list to verify that the configured nodes are connected to the fabric. Run the `ibhosts` command from a node that is linked to the fabric.

For example, log into `r1lead` and run the `ibhosts` command as follows:

```
r1lead:~ # ibhosts
Ca     : 0x0002c9030032bd50 ports 1 "r1i0n8 HCA-1"
Ca     : 0x0002c9030014a630 ports 1 "r1i0n7 HCA-1"
Ca     : 0x0002c9030014b140 ports 1 "r1i0n6 HCA-1"
Ca     : 0x0002c9030018cf00 ports 1 "r1i0n5 HCA-1"
Ca     : 0x0002c9030018cfa0 ports 1 "r1i0n13 HCA-1"
Ca     : 0x0002c9030018ce90 ports 1 "r1i0n15 HCA-1"
Ca     : 0x0002c9030014a610 ports 1 "r1i0n14 HCA-1"
Ca     : 0x0002c9030014b110 ports 1 "r1i0n16 HCA-1"
Ca     : 0x0002c9030018d170 ports 1 "r1i0n17 HCA-1"
```

The output shows each node connected, as expected, to the leader.

30. (Conditional) Increase the sweep interval to 90 seconds or more.

Complete this step if the cluster has more than 256 nodes.

Use a command such as the following:

```
# sgifmcli --set --id master-ib0 --arglist sweep-interval=90
```

Configuring Array Services for HPE Message Passing Interface (MPI) programs

You can configure compute nodes into an array. After you configure a set of compute nodes into an array, the Array Services software can perform authentication and coordination functions when HPE Message Passing Interface (MPI) programs are running. For more information, see the following:

HPE Message Passing Interface (MPI) User Guide

On a cluster with leader nodes, you cannot include the admin node or any leader nodes in an array. On a cluster without leader nodes, you can include the admin node in an array.

For general Array Services configuration information, see the manpages. The Array Services manpages reside on the admin node. If the HPE Message Passing Interface (MPI) software is installed on the admin node, you can retrieve the following manpages:

- `arrayconfig`(1M), which describes how to use the `arrayconfig` command to configure Array Services.
- `arrayconfig_smc`(1M), which describes Array Services configuration characteristics that are specific to clusters.

The procedures in the following topics assume the following:

- You want to create new a master image for the compute nodes.

  and

- You want to configure a new master image for the non-ICE compute nodes that you configured with user services.

After you create the preceding images, you can push out the new images to the compute nodes and to the compute services nodes.

The alternative is to configure Array Services directly on the nodes themselves. This method, however, leaves you with an Array Services configuration that is overwritten the next time someone pushes new software images to the cluster nodes.

NOTE: The procedures in the following topics assume that you want to install Array Services on a cluster with leader nodes. The steps you must complete for a cluster without leader nodes are similar. The major difference for clusters without leader nodes is that you do not need to perform steps that pertain to racks. On a cluster without leader nodes, do not perform the steps in the following procedures that pertain to racks and to leader nodes.

The following procedures explain how to configure Array Services.

- **Planning the configuration**
- **Preparing the Array Services images**
- **Power cycling the nodes and pushing out the new images**

Planning the configuration

The following procedure explains how to plan your array and how to select a security level.

Procedure

1. Log into the admin node as the root user.
2. Verify that the HPE MPI is installed on the cluster.

   If HPE MPI is not installed on the admin node, complete the following steps:

   - On RHEL 8 systems, enter the following command:

     ```
     # cm node dnf -n admin 'groupinstall HPE*MPI'
     ```

   - On RHEL 7 systems, enter the following command:

     ```
     # cm node yum -n admin 'groupinstall HPE*MPI'
     ```

   - On SLES systems, enter the following command:

     ```
     # cm node zypper -n admin 'groupinstall HPE*MPI'
     ```

3. Use the `cm node show` command to display a list of available nodes, and decide which nodes you want to include in the array.

   For example:

- To gather information about non-ICE compute nodes, enter the following command:

  ```
  # cm node show -t system compute
  ```

- To gather information about ICE compute nodes, enter the following command:

  ```
  # cm node show -t system ice_compute
  ```

The command shows all the compute nodes, including the compute nodes that might be configured as compute services nodes at this time.

4. Display a list of the available system images, and decide which images you want to edit.

For example, the following output is for an example cluster with leader nodes that is running in production mode:

```
# cm image show
lead-sles15sp1
lead-sles15sp1.prod1
sles15sp1
sles15sp1.prod1
ice-sles15sp1
ice-sles15sp1.prod1
```

The output includes image `sles15sp1.prod1`. The `sles15sp1.prod1` image is installed on a non-ICE compute node that is configured as a compute services node. Image `sles15sp1.prod1` is based on image `sles15sp1`, but it can include software to support user logins and a backup DNS server.

All system images are stored in the following directory:

```
/opt/clmgr/image/images
```

The preceding output shows the original, factory-shipped system images for the leader nodes, the non-ICE compute nodes, and the ICE compute nodes. These original files are as follows:

- `lead-sles15sp1`
- `sles15sp1`
- `ice-sles15sp1`

The output also shows customized images for this cluster with leader nodes. These file names end in `.prod1`, for production use, and are as follows:

- `lead-sles15sp1.prod1`
- `sles15sp1.prod1`. This image is the image that resides on the non-ICE compute services node.
- `ice-sles15sp1.prod1`

For each of these images, the associated kernel is `3.0.101-94-default`.

The examples in this Array Services configuration procedure add the Array Services information to the customized, production images with the `.prod1` suffix.

5. Decide what kind of security you want to enable.

Array Services includes its own authentication and security. If your site requires additional security, you can configure MUNGE security, which the installation includes. Your security choices are as follows:

- `munge` on all the nodes you want to include in the array. Configures additional security provided by MUNGE. The installation process installs MUNGE by default. If you decide to use MUNGE, the MPI from HPE configuration process explains how to enable MUNGE at the appropriate time.
- `none` on the compute services nodes and `none` on the compute nodes

  or

  `noremote` on the compute services nodes and `none` on the compute nodes

  These specifications have the following effects:

  - When you specify `none` on all the nodes you want to include in the array, all authentication is disabled.
  - When you specify the following, users must run their jobs directly from the compute services nodes:
    - `noremote` on the compute services nodes

      and
    - `none` on the compute nodes

    In this case, users cannot submit MPI from HPE jobs remotely.
- `simple` (default). Generates hostname/key pairs by using either the OpenSSL, `rand` command, 64-bit values (if available) or by using `$RANDOM` Bash facilities.

6. Proceed to the following:

   Preparing the Array Services images

Preparing the Array Services images

Before you create images that include Array Services, copy the production system images that your system is using now.

The following procedure explains how to prepare the images.

Procedure

1. Log into the admin node as the root user.
2. Use two `cm image copy` commands to clone the following:

   - One of the images that resides on a compute services node

     and
   - One of the images that resides on a compute node

   The format is as follows:

   ```
   cm image copy -s existing_image -i new_image
   ```

   For *existing_image*, specify the name of one of the existing images.

   For *new_image*, specify the new name for that to want to give to the image.

   For example, the following commands copy the first-generation production images to new, second-generation production images:

   ```
   # cm image copy -s sles15sp1.prod1 -i sles15sp1.prod2
   # cm image copy -s ice-sles15sp1.prod1 -i ice-sles15sp1.prod2
   ```

3. Enter the following command to change to the system images directory:

   ```
   # cd /opt/clmgr/image/images
   ```

4. (Optional) Use the `cp` command to copy the MUNGE key from the new compute services node image to the new compute node image.

   Complete this step if you want to configure the additional security that MUNGE provides.

   The MUNGE key resides in `/etc/munge/munge.key` and must be identical on all the nodes that you want to include in the array. The copy command is as follows:

   ```
   cp /opt/clmgr/image/images/new_computeservices_image/etc/munge/munge.key \
   /opt/clmgr/image/images/new_compute_image/etc/munge/munge.key
   ```

   For *new_computeservices_image*, specify the name of the new compute services node image you created.

   For *new_compute_image*, specify the name of the new compute node image you created.

   For example:

   ```
   # cp /opt/clmgr/image/images/sles15sp1.prod2/etc/munge/munge.key \
   /opt/clmgr/image/images/ice-sles15sp1.prod2/etc/munge/munge.key
   ```

5. Use the following command to install the new image on the compute services node:

   ```
   cm node provision -n hostname(s) -i new_computeservices_image -s
   ```

   For *hostname*, specify the hostname or hostnames of the compute services node. This node is the node that you want users to log into when they log into the array.

   For *new_computeservices_image*, specify the name of the new image you created.

   For example, the following command installs the new image on node `n1`:

   ```
   # cm node provision -n n1 -i sles15sp1.prod2 -s
   ```

6. Use the `ssh` command to log into the compute services node from which you expect users to run MPI from HPE programs.

   For example, log into `n1`.

7. Use the `arrayconfig` command to configure the compute services node and compute nodes into an array.

   You can specify more than one compute services node.

   The `arrayconfig` command creates the following files on the compute service node to which you are logged in:

   - `/etc/array/arrayd.conf`
   - `/etc/array/arrayd.auth`

   Enter the `arrayconfig` command in the following format:

   ```
   /usr/sbin/arrayconfig -a arrayname -f -m -A method node node ...
   ```

   For *arrayname*, specify a name for the array. The default is `default`.

   For *method*, specify `munge`, `none`, or `simple`. A later step explains how to specify `noremote` for a compute services node.

   For each *node*, specify a list of node IDs.

Example 1. To specify that array `myarray` use MUNGE security and include all ICE compute nodes, enter the following command:

```
# /usr/bin/arrayconfig -a myarray -f -m \
-A munge $(cnodes --compute --ice-compute)
```

Example 2. To specify that array `yourarray` use no security, include one compute service node, and include all ICE compute nodes, enter the following command:

```
# /usr/bin/arrayconfig -a yourarray -f -m \
-A none n0 $(cnodes --ice-compute)
```

8. Proceed to the following:

   Configuring the authentication files in the new system images on the admin node

Configuring the authentication files in the new system images on the admin node

Complete one of the following procedures, based upon whether you want to permit remote access to the compute services node:

- If you specified `–A munge` or `–A simple` for authentication

  or

  If you specified `–A none` for authentication, and you want to permit users to log into a compute services node remotely to submit MPI from HPE programs, proceed to the following:
  Permitting remote access to the compute services node

- If you specified `–A none` for authentication, and you want to prevent users from logging into a compute services node remotely to submit MPI from HPE programs, proceed to the following:

  Preventing remote access to the service node

Permitting remote access to the compute services node

The following procedure assumes that you want to permit job queries and commands on the compute services node. It explains how to copy the array daemon files to the admin node.

Procedure

1. Log into one of the compute services nodes as the root user.

2. Copy the `arrayd.auth` file and the `arrayd.conf` files from the compute services node to the new compute services node image on the admin node.

   Enter the following command:

   ```
   # scp /etc/array/arrayd.* \
   admin:/opt/clmgr/image/images/computeservices_image/etc/arrayd.*
   ```

   For *computeservices_image*, specify the compute services node image on the admin node.

   Enter this command all on one line. The command in this step uses a backslash (\) character to continue the command to the following line.

   For example:

   ```
   # scp /etc/array/arrayd.* \
   admin:/opt/clmgr/image/images/sles15sp1.prod2/etc/arrayd.*
   ```

3. Copy the `arrayd.auth` file and the `arrayd.conf` files from the compute services node to the new compute node image on the admin node.

Enter the following command:

```
# scp /etc/array/arrayd.* \
admin:/opt/clmgr/image/images/compute_image/etc/arrayd.*
```

For *compute_image*, specify the compute node image on the admin node. If the cluster has SU leader nodes, this is a non-ICE compute node image. If the cluster has ICE leader nodes, this is an ICE compute node image.

Enter this command all on one line. Notice that the command in this step uses a backslash (\) character to continue the command to the following line.

For example:

```
# scp /etc/array/arrayd.* \
admin:/opt/clmgr/image/images/ice-sles15sp1.prod2/etc/arrayd.*
```

4. Proceed to the following:

   Distributing images to all the nodes in the array

Preventing remote access to the service node

You can prevent a compute services node from receiving any requests from other computers on the network. In this case, the compute services node can send requests to all remote nodes, but it does not listen on TCP port 5434 for any incoming requests. Complete the procedure in this topic if this behavior is required at your site.

The following procedure explains how to accomplish the following:

- How to configure the Array Services files to prevent remote access
- How to copy the array daemon files to the admin node

Procedure

1. Log into one of the compute services nodes as the root user.
2. Open the following file with a text editor:

   `/etc/array/arrayd.auth`
3. Enter the following, all on one line:

   ```
   AUTHENTICATION NOREMOTE
   ```
4. Save and close the file.

   Make sure that the file contains only the one line.
5. Enter the following command to copy `/etc/array/arrayd.auth` and `/etc/array/arrayd.conf` from the compute services node to the new compute services node image on the admin node:

   ```
   # scp /etc/array/arrayd.* \
   admin:/opt/clmgr/image/images/computeservices_image/etc/arrayd.*
   ```

   For example:

   ```
   # scp /etc/array/arrayd.* \
   admin:/opt/clmgr/image/images/sles15sp1.prod2/etc/arrayd.*
   ```
6. Log into the admin node as the root user.
7. Create file `/opt/clmgr/image/images/`*`compute_image`*`/etc/array/arrayd.auth`.

For example:

```
# vi /opt/clmgr/image/images/sles15sp1.prod2/etc/array/arrayd.auth
```

8. Add the following all on one line:

   ```
   AUTHENTICATION NONE
   ```

9. Save and close the file.

   Make sure that the file contains only the one line.

10. Enter the following command to copy the `/etc/array/arrayd.conf` file to the compute nodes:

    ```
    # scp /etc/array/arrayd.conf \
    admin:/opt/clmgr/image/images/compute_image/etc/arrayd.conf
    ```

    For example:

    ```
    # scp /etc/array/arrayd.conf \
    admin:/opt/clmgr/image/images/ice-sles15sp1.prod2/etc/arrayd.conf
    ```

11. Back up the images.

    ```
    # cm image revision commit -i sles15sp1.prod2 \
    -m "configured array services"
    # cm image revision commit -i ice-sles15sp1.prod2 \
    -m "configured array services"
    ```

12. Proceed to the following:

    Distributing images to all the nodes in the array

Distributing images to all the nodes in the array

The following procedure explains how to complete the following tasks:

- Assign the new compute services node image to the compute services nodes
- Assign the new compute node image to the compute nodes

Procedure

1. Log into the admin node as the root user.
2. Assign the new compute services node image to the compute service nodes.

   Use one or more `cm node provision` commands in the following format:

   ```
   cm node provision -n hostname -i new_computeservice_image -s
   ```

Variable	Specification
hostname	Specify the hostname for one or more of the compute services nodes. In the cluster definition file, this name is the name that appears in the `hostname1` field. You can specify the `*` wildcard character to represent a string of identical characters in this field. Use wildcard characters in the following situation: • If you want to specify more than one hostname and • If your nodes have names that are similar For example, if your hostnames are `n1`, `n2`, `n3`, and `n57`, specify `n*` in this field if you want to specify all compute services nodes.
new_service_image	Specify the name of the new compute services node image you created.

Example 1. The following command assigns the new compute services node image to all compute services nodes:

```
# cm node provision -n n* -i sles15sp1.prod2 -s
```

Example 2. The following command assigns the new compute services node image to `n101`:

```
# cm node provision -n n101 -i sles15sp1.prod2 -s
```

3. Proceed to the following:

   Power cycling the nodes and pushing out the new images

Power cycling the nodes and pushing out the new images

Procedure

1. Enter the following command to reboot the compute services nodes and the compute nodes:

   ```
   # cm power reboot -t node "*"
   ```

2. Use one or more `cm power` commands in the following format to power off the compute nodes that you want to reimage:

   ```
   cm power off -t node "hostname"
   ```

   For *hostname*, specify one or more compute node hostnames.

   If you have many compute nodes, you can use wildcard characters.

   For example, the following command powers off all the compute nodes on a cluster with leader nodes:

   ```
   # cm power off -t node "r*i*n*"
   ```

   Issue as many `cm power` commands as needed.

3. (Conditional) Assign the new compute node image to the compute nodes.

   Complete this step if the cluster has non-ICE compute nodes that use an NFS root file system.

   The following tables contain the instructions you need to complete this step.

   Scenario 1 - Complete the following steps on clusters with non-ICE compute nodes that have NFS root file systems.

Step	Task

a.	Run the `activate-nfs-image` command in the following format to activate the NFS image:
	`activate-nfs-image new_compute_image`
	For *new_compute_image*, specify the name of the new compute node image you created.
	For example:
	`# activate-nfs-image sles15sp1.prod2`
	`Activate image - Syncing image to RO NFS path`
	`.`
	`.`
	`.`
	On clusters with SU leader nodes, this command can take a few moments to complete.

b.	Assign the new non-ICE compute node images to the non-ICE compute nodes. For example, use the following `cimage` command: `# cimage --set sles15sp1.prod2 3.0.101-108.38-default "r*i*n*"`

Scenario 2 - Complete the following steps on clusters with ICE leader nodes:

Step	Task
a.	Ensure that the node(s) are fully booted.

b.	Run the `cimage` command in the following format:
	`cimage --push-rack new_compute_image rack`
	For *new_compute_image*, specify the name of the new ICE compute node image you created.
	For *rack*, specify the IDs of the racks in which the ICE compute nodes reside.
	For example, the following command pushes the ICE compute node image to all ICE compute nodes in all racks:
	`# cimage --push-rack ice-sles15sp1.prod2 r*`

c.	Assign the new ICE compute node image to the ICE compute nodes. For example, use the following `cimage` command on an HPE SGI 8600 cluster: `# cimage --set ice-sles15sp1.prod2 3.0.101-108.38-default "r*i*n*"`

4. Use one or more `cm power` commands in the following format to start the compute nodes:

```
cm power on -t node "hostname"
```

For *hostname*, specify the hostnames of the compute nodes.

If you have many compute nodes, you can use wildcard characters.

For example, the following command powers on all compute nodes:

```
# cm power on -t node "r*i*n*"
```

Issue as many `cm power on` commands as needed.

Configuring ICE compute nodes to use a non-ICE compute node as a network address (NAT) gateway (HPE SGI 8600 clusters)

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 non-ICE compute node.

Procedure

1. Use the documentation from your software distribution to configure a NAT gateway on a non-ICE compute node.

   During the configuration, update the static routes.

2. 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 non-ICE compute node.

3. 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.

4. 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 chassis, and the slot of the ICE compute node. Some examples are available in the script.

5. Push the new image, as follows:

- Enter the following command to shut down and stop all the ICE compute nodes:

  ```
  admin# cm power halt -t node "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# cm power on -t node "r*i*n*"
  ```

When you power up the ICE compute nodes, the new default gateway route is present on all ICE compute nodes.

Troubleshooting cluster manager installations

Troubleshooting configuration changes

If a configuration change does not affect the cluster in the intended manner, try one of the following approaches:

- Edit the node image on the admin node. For example, you can try the following:

  1. On the admin node, reconfigure the image for the compute nodes that you use for user services
  2. Reimage the compute services nodes with the new, reconfigured image.

- Edit the node customization scripts.

  For example, the ICE compute node update scripts reside on the admin node in the following directory:

  ```
  /opt/sgi/share/per-host-customization/global
  ```

Verifying the switch cabling

If the switches are not working, the first troubleshooting step is to verify the switch cabling.

The following figure shows a switch stack with two switches. In this switch stack, the two switches constitute the spine switch stack. One is the master switch and the other is the slave switch.

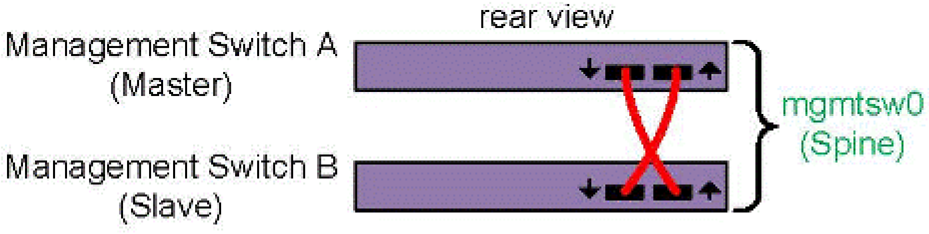


Figure 6: Spine switch stack with two switches

The following figure shows a switch stack with multiple switches. The first two switches constitute the spine switch stack, and the other switches constitute the secondary switch stack.

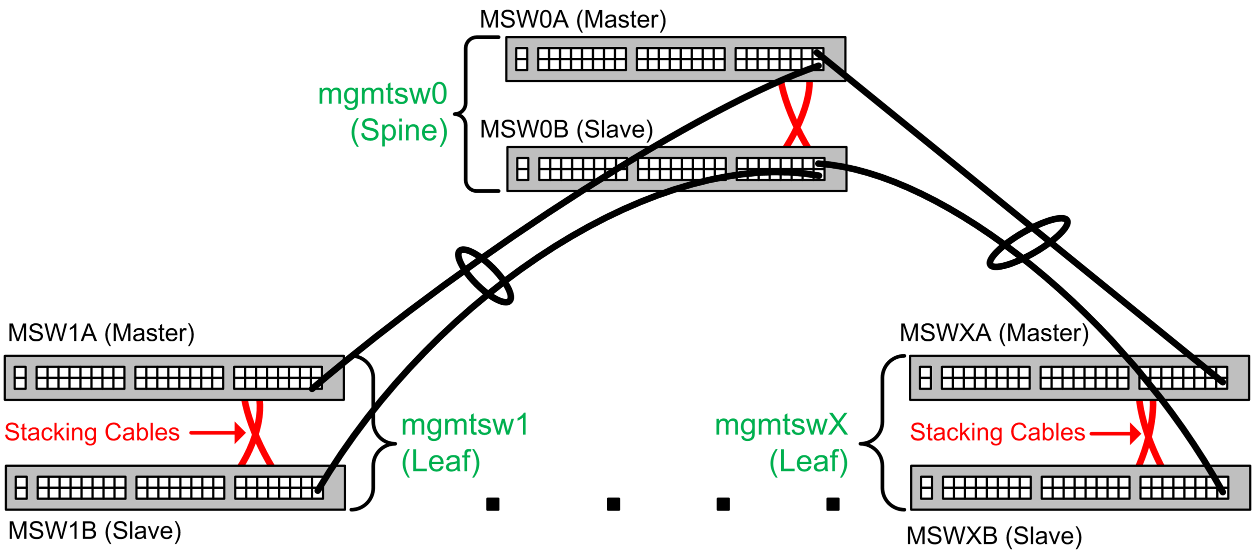


Figure 7: Switch stack with multiple switches

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

Procedure

1. Visually inspect your system.

   Note the types of switches you have and their identifiers. At a minimum, you have at least one stacked (or non-stacked) management switch. The management switch that is connected directly to admin node is almost always considered the **spine** switch. Additional stacked or non-stacked switches connect to the spine switch. These additional switches are almost always considered to be **leaf** switches. In some configurations, leaf switches can connect to other leaf switches, and this creates a **multi-tiered** management network topology.

   When multiple physical switches are in a stacked configuration, those multiple physical switches can be thought of as a single **logical** switch. This means that the logical switch is assigned one IP address for remote management, and the `switchconfig` command can to configure the entire switch stack.

   NOTE: Determine whether your system contains management switches from Arista Networks, Inc. and whether the switches are using Multi-Chassis Link Aggregation (MLAG). In this case, each switch in an MLAG pair is independent and cannot be considered a stacked switch. Each switch in an MLAG pair receives an IP address separately and is managed separately.

2. Determine whether or not you have a cluster definition file.

   If you have a cluster definition file that contains the MAC addresses of the cluster components, you can safely have all nodes and all switches powered on when you run the `discover` command. During the `discover` process, a cluster definition file ensures that a node with a MAC address is assigned an IP address that matches the node MAC address.

   If you do not have a cluster definition file, all nodes and switches other than the admin node must begin in a powered off state. Without a cluster definition file, the process involves powering on one node at a time and waiting for a DHCP request from the newly powered on node to reach the admin node. The admin node associates the MAC address from this new DHCP request with the node being discovered.

3. (Conditional) Disconnect the secondary, redundant cables that connect switches together.

   Complete this step if you have not yet run the `discover` command to configure the management switches. Or, complete this step if you plan to reset the management switches back to factory default settings. This action prevents networking loops.

Use Method 1 or Method 2 to disconnect the switches. The instructions for both methods include an example that assumes that mgmtsw0 is connected to mgmtsw1 with the following cable mappings:

- mgmtsw0 port 1/48 ---- mgmtsw1 port 1/48
- mgmtsw0 port 2/48 ---- mgmtsw1 port 2/48

Also assume that mgmtsw1 needs to be reset to factory settings. The reset could be needed to obtain a fresh configuration, to update the VLANs or IP addresses on mgmtsw1, or for any reason.

Method 1 - Software method

If the spine switch is reachable from the admin node, you can prevent a networking loop when mgmtsw1 is factory reset. From the admin node, complete the following steps:

a. Enter the following command to disable the redundant port that connects mgmtsw0 to mgmtsw1

```
# switchconfig port -s mgmtsw0 --disable -p 2/48
```

b. Enter the following command to reset mgmtsw1 back to factory default settings:

```
# switchconfig reset_factory_defaults -s mgmtsw1 --force
```

c. Wait 3~10 minutes for mgmtsw1 to come back online. Enter the following command:

```
# ping mgmtsw1
```

d. Enter the following command to reconfigure **mgmtsw1**:

```
# switchconfig_configure_node --node mgmtsw1
```

Wait for this command to complete.

e. After mgmtsw1 is configured correctly, enter the following command to re-enable the redundant port that you disabled earlier in this procedure.

```
# switchconfig port -s mgmtsw0 --enable -p 2/48
```

f. (Conditional) Reapply lost configuration attributes.

Complete this step if, for example, mgmtsw1 had any nodes that require a switch configuration that was lost in this procedure. For example, these nodes might be a scalable unit (SU) leader node, an ICE leader node, or compute nodes that use 802.3ad (LACP) bonding.

To reapply any lost configuration settings, use commands such as the following:

```
# switchconfig_configure_node --node r1lead
# switchconfig_configure_node --node leader1
# switchconfig_configure_node --node service100
# switchconfig_configure_node --node r1lead,service100
```

Method 2 - Manual method

If you need to reset all management switches on your cluster or have lost full connectivity to the management fabric, you need physical access to the cluster hardware. This method, Method 2, is the same as Method 1, but the initial step is different. Rather than use the switchconfig command to disable ports, start the procedure by doing one of the following to replace Step a:

- Unplug all redundant switch cabling from one end of the wire for all cabling between management switches and for all cabling between management switches and chassis management controllers (CMCs).

  OR

- Attach a serial connection to a management switch, open up a serial console session, and use the vendor-specific methods to temporarily disable the redundant ports until switches can be successfully configured again.

The following figure shows an example topology with 3 management switches (1 spine switch stack and 2 leaf switch stacks) and which cables to disconnect.

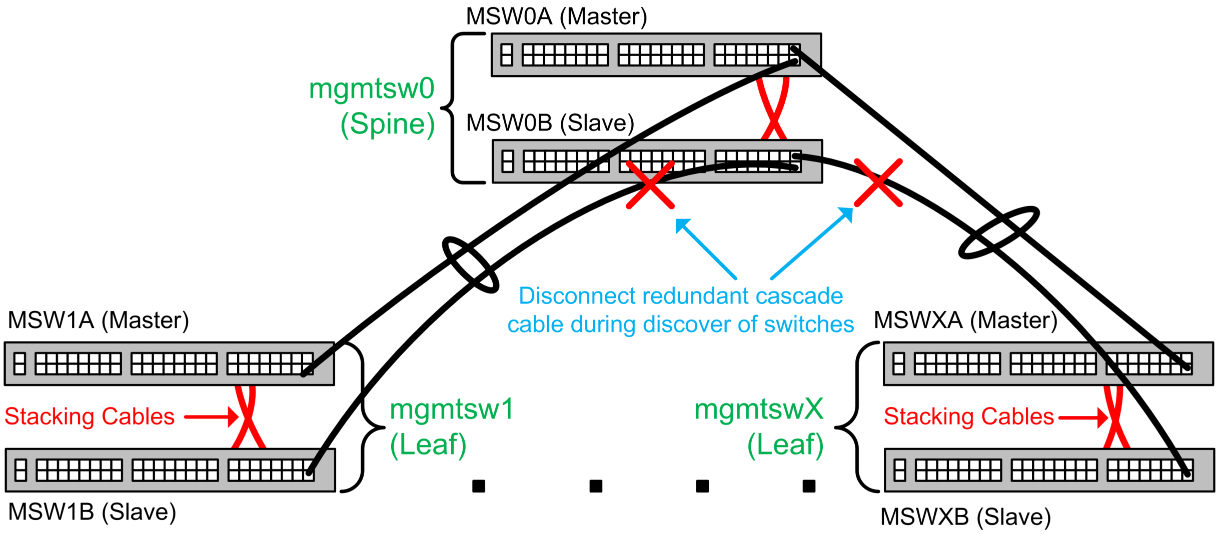


Figure 8: Cables to disconnect

Chassis management controllers (CMCs) failed to configure

The following topics provide background information about CMCs:

- **About the `cmcdetectd` service**
- **About chassis management controllers (CMCs) and VLANs on HPE Apollo 9000 clusters**
- **About chassis management controllers (CMCs) and VLANs on HPE SGI ICE clusters**

If you suspect that the CMCs failed to configure automatically, look in the following log file for information regarding the status of CMCs in the system:

```
/var/log/cmcdetectd.log
```

If you find errors in the preceding log files, power on the CMCs and use one of the following procedures to configure the CMCs:

- **Reviewing the chassis management controller (CMC) configuration**
- **Method 1 - Configuring the chassis management controller (CMC) switches manually**
- **Method 2 - Configuring the chassis management controller (CMC) switches manually**

About the cmcdetectd service

The cmcdetectd service runs on the admin node as a daemon. This service uses a specific tcpdump command to listen to CMC-generated DHCP packets on the management network. When the cmcdetectd service receives a CMC-generated DHCP packet, it takes the following actions:

- It inspects the information located in the packet.
- It determines the appropriate VLAN and bonding settings to apply to the attached management switch ports connected to the CMC in question.

CMCs are cabled in the same manner as other redundantly cabled components in the cluster. For example, assume that Rack 1, CMC 0, Port 1 is connected to mgmtsw0 port 1/11. In this case, Rack 1, CMC 0, Port 2 must be connected to mgmtsw0 port 2/11, and so on.

About chassis management controllers (CMCs) and VLANs on HPE Apollo 9000 clusters

The HPE Apollo 9000 system CMC default values are as follows:

- The default untagged VLAN numbering begins at 2001. The end number depends on how many CMCs are allowed in a management VLAN. By default, 8 CMCs are allowed per management VLAN. By default, the VLAN begins at VLAN 2001 and ends at VLAN 2999.
- The HPE Apollo 9000 CMCs do not used a tagged cooling VLAN.

The following commands retrieve VLAN information or configure VLAN settings:

- To view how many CMCs are allowed in a management VLAN, use the following command:

  ```
  # cadmin --show-cmcs-per-mgmt-vlan
  8
  ```

- The number of CMCs allowed in a management VLAN can be one of the following:

  - 0.

    NOTE: On Apollo 9000 systems, when *number* is set to 0, it disables the automatic generation of VLANs for CMCs.

  - An integer that is a multiple of 4 and no greater than 48.

  To change how many CMCs are allowed in a management VLAN, use one of the following methods.

  Method 1. Use the following command:

  ```
  # cadmin --set-cmcs-per-mgmt-vlan number
  ```

  For *number*, specify an integer that is a multiple of 4 and is no greater than 48.

  Method 2. Use the cluster configuration tool, as follows:

  1. Enter the following command on the admin node:

     ```
     # configure-cluster
     ```

  2. Click through **Initial Setup Menu > Network Settings > Configure Management Network VLAN Settings > CMCs per Management VLAN.** Specify the setting.

- To view the management VLAN start value and end value, use the following commands:

  ```
  # cadmin --show-mgmt-vlan-start
  2001
  # cadmin --show-mgmt-vlan-end
  2999
  ```

- To change the management VLAN start value and end value, use the following commands:

  ```
  cadmin --set-mgmt-vlan-start number
  cadmin --set-mgmt-vlan-end number
  ```

About chassis management controllers (CMCs) and VLANs on HPE SGI ICE clusters

The default untagged VLAN numbering begins at *rack_number* + 100. By default, the untagged rack VLAN numbering begins at VLAN 101 and ends at VLAN 1100.

The tagged VLAN is the cooling VLAN number, which is the MCell network. By default, the tagged cooling VLAN is set to VLAN 3. This means that the first rack of ICE components, rack 1, is assigned to VLAN number of 101. Rack 2 is assigned to VLAN number 102 and so on. To view these values, use the following commands:

-
  ```
  # cattr get rack_vlan_start
  101
  ```

-
  ```
  # cattr get rack_vlan_end
  1100
  ```

-
  ```
  # cattr get mcell_vlan
  3
  ```

To change the default VLAN values, use the following commands:

- `cattr set rack_vlan_start number`
- `cattr set rack_vlan_end number`
- `cattr set mcell_vlan number`

NOTE: If you change these values, run the `discover` command again to reconfigure all rack leader controllers (RLCs) and their associated ICE compute blades.

Reviewing the chassis management controller (CMC) configuration

The following procedure explains how to review the current CMC configuration.

Procedure

1. From a local console, use the `ssh` command to open a remote session to the admin node.
2. Enter the following command to ensure that the `cmcdetectd` service is running:

   ```
   # systemctl status cmcdetectd
   ```

   If the `cmcdetectd` service is not running, enter the following command:

   ```
   # systemctl start cmcdetectd
   ```

3. Enter the following command to monitor the progress of cmcdetectd:

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

4. (Optional) Reset the management switch ports.

   If you know the management switch and the ports to which a CMC connects, you can reset the management switch ports back to default settings.

   This practice ensures that the cmcdetectd service receives the DHCP packets from the CMC.

   Example: Rack 1, CMC 0 is plugged into mgmtsw0 ports 1/11 and 2/11. Issue the following command to set both ports back to default settings:

   ```
   # switchconfig unset --switches mgmtsw0 --ports 1/11 --redundant yes
   ```

5. Flip the power breakers on for the CMCs, one rack at a time.

   Notice that cmcdetectd runs in a serial fashion. It handles one CMC configuration at a time to prevent configuration conflicts on the management switches.

6. Verify the configuration.

   After the cmcdetectd service configures a CMC, the system logs an entry to the following file:

   /etc/cmc-switch-info.txt

   Verify that this file contains the correct entries.

   Example 1. For an HPE Apollo 9000 cluster, the CMC entry might look as follows:

   ```
   # cat /etc/cmc-switch-info.txt
   mac_address=00:fd:45:ff:3b:46, mgmtsw=mgmtsw1, vlans=None, default_vlan=2001, bonding=manual, ports=1/0/1,
   redundant=yes, cmc_type=nonice, cmc_hostname=r1c0
   # VLAN to management switch configuration
   vlan=2001, mgmtsw=mgmtsw1, configured=yes, vlan_type=nonice
   ```

   Example 2. For an ICE cluster, the CMC entry might look as follows:

   ```
   # cat /etc/cmc-switch-info.txt
   mac_address=08:00:69:15:ce:38, mgmtsw=mgmtsw0, vlans=3, default_vlan=101, bonding=manual, ports=1/0/11,
   redundant=yes, cmc_type=ice, cmc_hostname=None
   # VLAN to management switch configuration
   vlan=101, mgmtsw=mgmtsw0, configured=no, vlan_type=ice
   ```

7. Back up the /etc/cmc-switch-info.txt file to another server at your site.

   If you ever have to perform a disaster recovery, this information can be useful.

Method 1 - Configuring the chassis management controller (CMC) switches manually

To reapply the CMC configuration on management switches quickly, complete this procedure. Use this procedure when the management switches are reset to factory settings or when a switch is replaced. Use this procedure if you have the /etc/cmc-switch-info.txt file.

Procedure

1. Make sure that all CMCs are configured.

2. Run the following command to configure the CMC switches:

   ```
   # cmcdetectd --switchconfig
   reading the CMC-Switch configuration file: /etc/cmc-switch-info.txt ...

   configuring {'mgmtsw0': '172.23.255.254'}...
   ```

```
CMC mac-address = ['08:00:69:15:ce:38']
switchport(s) = ['1:11']
native vlan = 101
tagged vlan(s) = ['3']
bonding mode = manual
redundant = yes

...result was successful!

switchconfig functions completed successfully for all CMC's - ["['08:00:69:15:ce:38']"]
to save configuration, please use `switchconfig config --save --switches <switch hostname>
```

Method 2 - Configuring the chassis management controller (CMC) switches manually

Use this procedure if you do not have the `/etc/cmc-switch-info.txt` file.

This manual configuration method requires that you provide the following information:

- The rack VLANs in which the CMCs reside.
- The physical ports and management switches to which the CMCs are cabled.

If you cannot provide this information, do not use this method to configure the CMCs.

The following procedure explains how to configure the CMCs manually.

Procedure

1. From a local console, use the `ssh` command to open two remote sessions to the admin node.
2. In one remote session window, enter the following command to monitor the `switchconfig` log file:

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

   Your goal is to make sure that the commands being sent are completing successfully.
3. In the other remote session window, use a `switchconfig` command to configure the management switch manually.

   As inputs to the `switchconfig` command, use the rack VLAN information and information about the physical port location of the CMC.

   For example, assume that CMC `r1i0c` (that is, rack 1 CMC0 using VLAN 101) is connected to management switch `mgmtsw0` on port `1:11` and `mgmtsw0` port `2:11`. You can use one of the following commands:

   ```
   # switchconfig set --switches mgmtsw0 --default-vlan 101 --vlan 3 \
   --bonding manual --ports 1:11 --redundant yes
   ```

   or

   ```
   # switchconfig set -s mgmtsw0 -d 101 -v 3 -b manual -p 1:11 -r yes
   ```

   For more information about the `switchconfig` command, enter the following:

   ```
   # switchconfig set --help
   ```
4. Flip the power breakers on for the CMCs, one rack at a time.

   Because the management switch is already configured, no additional tasks should be needed. When configured correctly, each leader node detects its CMCs shortly after the CMCs are powered on.
5. (Optional) Validate the CMC configuration.

Enter the following `switchconfig` command to display the bonding and VLAN configuration of the CMCs that are connected to the management network:

```
# switchconfig sanity_check -s mgmtsw0

======== Beginning Sanity Check on mgmtsw0 ========
.
.
.
checking port-channel sharing configuration on mgmtsw0...(address-based L2/L3/L3_L4 = static port-channel,
address-based L2/L3/L3_L4 lacp = LACP port-channel)

        port-channel group master is 1:17 with the following ports in a port-channel: 1:17, 2:17 in
        bonding mode: address-based L2
        port-channel group master is 1:18 with the following ports in a port-channel: 1:18, 2:18 in
        bonding mode: address-based L2
        port-channel group master is 1:19 with the following ports in a port-channel: 1:19, 2:19 in
        bonding mode: address-based L2
        port-channel group master is 1:20 with the following ports in a port-channel: 1:20, 2:20 in
        bonding mode: address-based L2
        port-channel group master is 1:21 with the following ports in a port-channel: 1:21, 2:21 in
        bonding mode: address-based L2
        port-channel group master is 1:22 with the following ports in a port-channel: 1:22, 2:22 in
        bonding mode: address-based L2
        port-channel group master is 1:23 with the following ports in a port-channel: 1:23, 2:23 in
        bonding mode: address-based L2
        port-channel group master is 1:24 with the following ports in a port-channel: 1:24, 2:24 in
        bonding mode: address-based L2
        port-channel group master is 1:3 with the following ports in a port-channel: 1:3, 2:3 in
        bonding mode: address-based L2 lacp

checking port-channel VLAN configuration on mgmtsw0 (Native VLAN means untagged packets are put into the VLAN,
Tagged VLAN means the port allows 802.1Q tagged packets to pass on the port)

        ======== General Information about VLAN Configurations (Correct Standard Configuration) ========

        Admin Node - Native(untagged) VLAN 1, allowed tagged VLANs: 3
        Service Node - Native(untagged) VLAN 1, allowed tagged VLANs: None
        Rack Leader Controller - Native(untagged) VLAN 1, allowed tagged VLANs:  (101, 102, etc.)
        CMC's - Native(untagged) VLAN  (101, 102, etc.), allowed tagged VLANs: 3
        Node BMC's - Native(untagged) VLAN 1, allowed tagged VLANs: None
        Cooling Equipment (CDU/CRC's) - Native(untagged) VLAN 3, allowed tagged VLANs: None
        Interswitch Links - Native(untagged) VLAN 1, allowed tagged VLANs: 3, 1999

        port-channel 1:17 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:18 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:19 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:20 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:21 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:22 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:23 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:24 is in Native(untagged) VLAN 101 and has the following allowed tagged VLANs: 3
        port-channel 1:3 is in Native(untagged) VLAN 1 and has the following allowed tagged VLANs: 101

======== Sanity Check Finished on mgmtsw0 , check above output for General Tips, Troubleshooting, and Results ========
```

Node provisioning takes too long or fails to complete

Symptom

Node provisioning (or imaging) takes too long or fails to complete.

Cause

By default, the cluster manager uses UDPcast to install software on each node. For some clusters, using the BitTorrent or `rsync` file transfer method helps the `discover` command to complete more quickly.

Action

1. Review the information in this step and select a different file transport method.

When you use the `discover` command to configure the cluster components, consider the following:

- The types of nodes you have
- The file transfer methods
- Whether you want to modify the node characteristics that currently exist in the cluster definition file

The file transfer method directly affects the time it takes to install software on each node. This process includes the following event sequence:

- The `discover` command completes and returns you to the system prompt.
- The leader nodes and non-ICE compute nodes install themselves with software from the admin node.

  On a cluster with leader nodes, the admin node pushes the compute node software to the leader nodes. The compute nodes install themselves with software from the leader node.
- The node comes up. At this point, if you issue a `cm power` command query to the node, the node responds with `ON`.

  For example, assume that you want to make sure that some ICE compute nodes are installed. Use the following command to verify the software installation on rack 2:

  ```
  # cm power status -t node "r2i*n*"
  r2i1n0      ON
  r2i1n17     ON
  ```

Regardless of the cluster type or node type, the default transfer method is UDPcast. Other file transfer methods are `rsync` and BitTorrent.

Table 2: Compute node characteristics and image information shows the node types and includes information about file transfer methods that are appropriate for each node.

Table 2: Compute node characteristics and image information

	Compute nodes with root disks	Non-ICE or ICE compute nodes with NFS root file systems	Nodes with `tmpfs` root file systems
Transport path	The admin node installs the flat compute nodes. Uses UDPcast, BitTorrent, or `rsync`.	Each node uses NFS to mount a root file system from its leader node.	On clusters without leader nodes, the admin node installs the nodes using UDPcast, BitTorrent, or `rsync`. On clusters with leader nodes, the leader nodes can aid in the provisioning process.
Software to be installed	The image resides on the admin node.	On clusters with leader nodes, the leader nodes provide the root file system using NFS.	On clusters with leader nodes, the leader nodes can transfer the image.

Table Continued

	Compute nodes with root disks	Non-ICE or ICE compute nodes with NFS root file systems	Nodes with `tmpfs` root file systems
Boot persistent?	Yes.	For ICE compute nodes, yes. For non-ICE compute nodes, boot persistence is possible depending on the configuration.	No. All is lost on reboot.
Node image memory use	No.	For ICE compute nodes, no. For non-ICE compute nodes, node image memory use depends on the configuration.	Yes. The root file system consumes system memory.
RPM installation notes	N/A	For ICE compute nodes, you cannot install RPMs on the nodes because the root file system is NFS read-only. For non-ICE compute nodes, NFS solutions that use overlay let RPMs be installed.	You can install RPMs on the nodes. However, each node receives a new image when the node boots, so RPM images are not boot-persistent.
Image root file system notes	N/A	For ICE compute nodes, the NFS root file system is read-only. Some locations are read/write. For non-ICE compute nodes, the overlay solutions are writeable. The overmount solutions are writeable to some locations.	N/A

The other consideration when choosing a file transport method is the method itself. **Table 3: File transfer methods** shows the available file transport methods.

Table 3: File transfer methods

	`rsync`	BitTorrent	UDPcast
Status	Not default	Not default	Default
Performance	Slower performance when pushing images to more than two nodes simultaneously.	Midrange performance.	Fastest performance.
Method	Pushes the image to all nodes, in separate sessions, over the cluster network. This action can consume all bandwidth when more than two nodes are involved.	Transfers the node image as a `tar` file that is divided into pieces. The individual nodes receive the pieces and assemble the pieces into an image. After the node assembles the image, it boots. When you use this method, the miniroot is always transferred using `rsync`. The other image components are transferred using BitTorrent.	Transfers the node image in a multicast stream, which has one sender and many listeners.
Encryption?	Yes	No	Yes
Appropriateness	Suited for a small number of nodes (2-5). If you have many nodes, run the `discover` command multiple times on groups of nodes each time.	Suited for a large number of nodes.	Most efficient for large numbers of nodes. Requires the switches to be configured for multicast traffic. Some switches might require additional configuration. Switches shipped with HPE clusters require no additional configuration.
Files transferred	Kernels, `initrd`, and the miniroot file system.	Kernels, and `initrd`. The system uses `rsync` to transfer the miniroot file system.	The miniroot file system.

In addition to the tables, you can use the following figures to help you select a transport method:

- **Figure 9: Selecting a transport method for provisioning clusters that have leader nodes**
- **Figure 10: Selecting a transport method for provisioning clusters that do not have leader nodes**

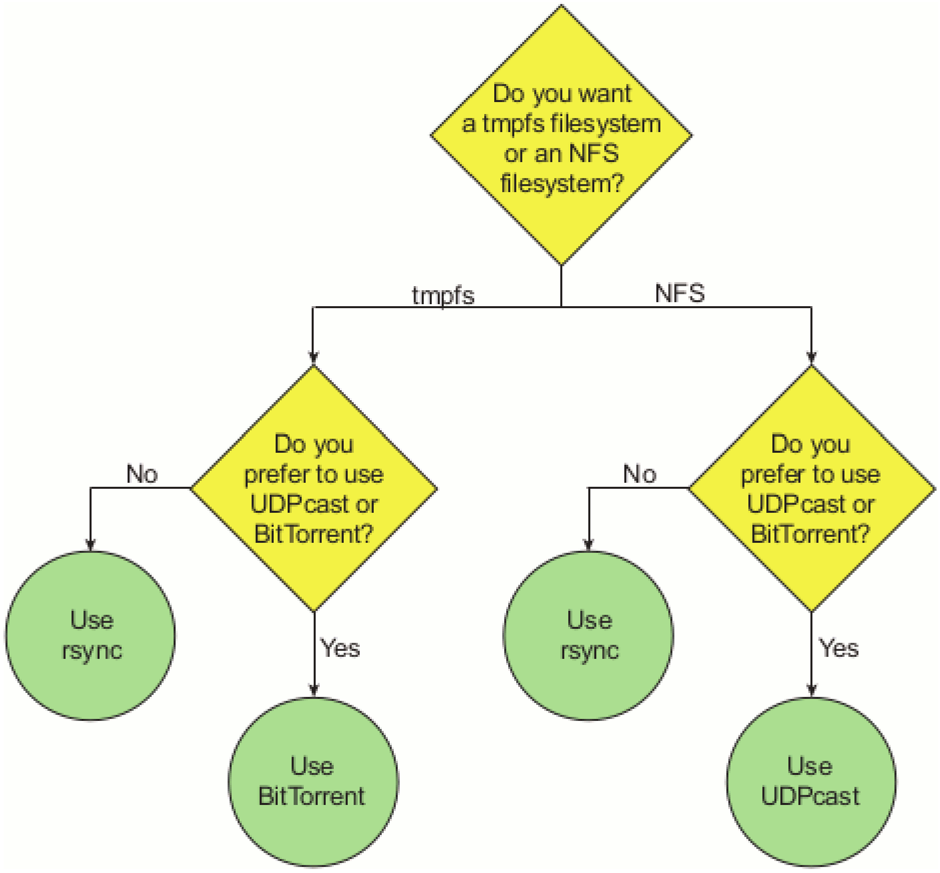

Do you want a tmpfs filesystem or an NFS filesystem?
tmpfs
NFS
Do you prefer to use UDPcast or BitTorrent?
No
Use rsync
Yes
Use BitTorrent
Do you prefer to use UDPcast or BitTorrent?
No
Use rsync
Yes
Use UDPcast

Figure 9: Selecting a transport method for provisioning clusters that have leader nodes

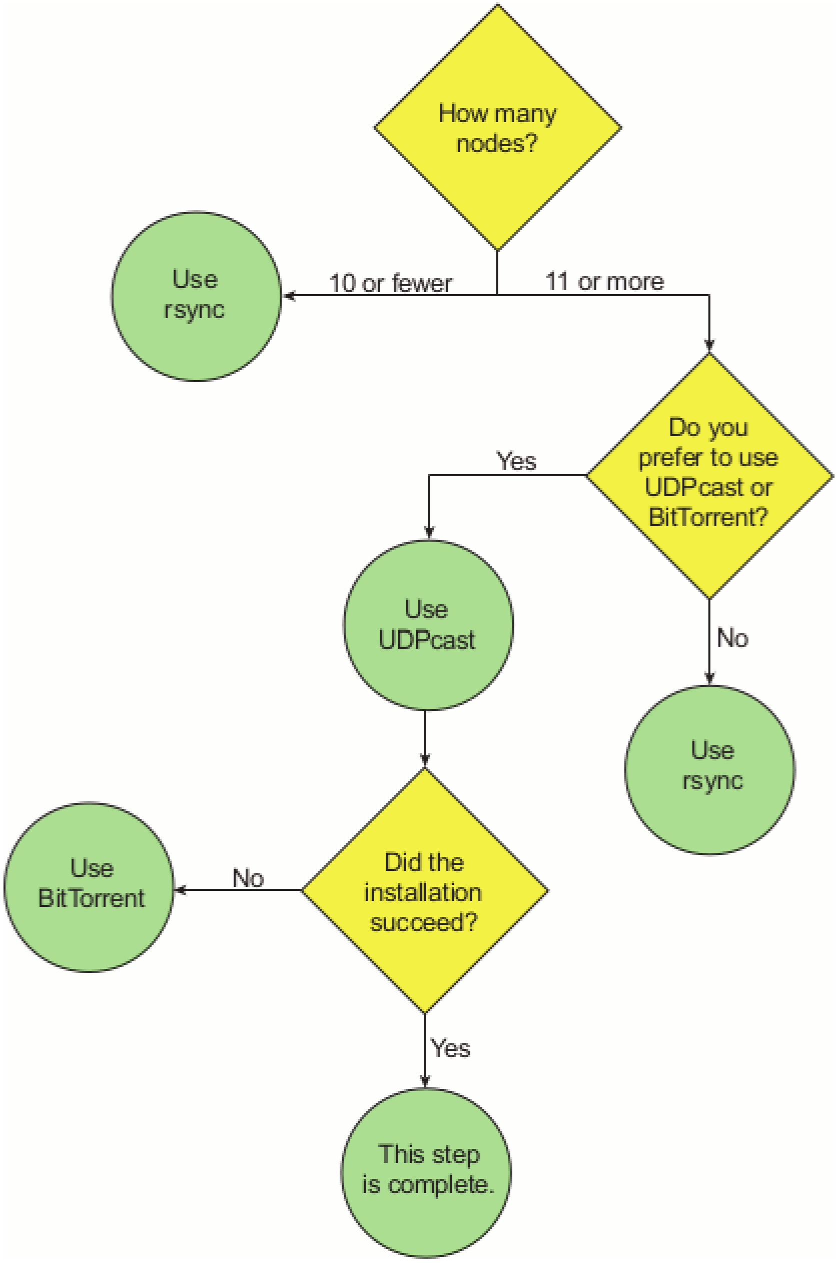


Figure 10: Selecting a transport method for provisioning clusters that do not have leader nodes

2. Update the cluster definition file to specify the alternative file transport method.

   Specify one of the following settings:

   - `transport=udpcast`
   - `transport=bt`
   - `transport=rsync`

For example, to change the transport method for ICE leader node r1lead to be UDPcast, add the setting to the end of the r1lead specification, as follows:

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

3. Run the discover command again, and specify the newly updated cluster definition file.
4. Evaluate the provisioning time with the new cluster definition file.

   In some cases, any file transfer method can result in compute nodes or leader nodes that do not complete the data transfer. If the data transfer does not finish, reinstall the software on the failed node.

   Use the cm node provision command to install the software:

   ```
   cm node provision -n failed_node_name
   ```

   If you continue to have problems, you might have to change your transport method again. If UDPcast and BitTorrent both fail, specify rsync. Your site network configuration can affect the speed at which the discover command can push software to nodes.

Suppressing nonfatal messages in the authentication agent

Symptom

The system issues the following erroneous message when you use ssh to log into the admin node as the root user:

```
Could not open a connection to your authentication agent.
```

You can safely ignore this message. Alternatively, use the command in this topic to start the agent in a way that suppresses this message.

Action

Start the ssh agent in a way that suppresses nonfatal messages about the authentication agent.

For example, in the bash shell, enter the following command:

```
# exec ssh-agent bash
```

To suppress the erroneous message, run this command after each boot during the installation process. After installation, the system no longer issues this message.

Verifying that the clmgr-power daemon is running

The following procedure explains how to make sure that the clmgr-power daemon is running properly.

Procedure

1. Log into the admin node as the root user, and enter the following command to make sure that the clmgr-power daemon is running:

   ```
   # systemctl status clmgr-power
   ```

For example, the following example shows the daemon running as expected on a SLES system:

```
# systemctl status clmgr-power
● clmgr-power.service - clmgr power
   Loaded: loaded (/usr/lib/systemd/system/clmgr-power.service; enabled;
vendor preset: enabled)
   Active: active (exited) since Tue 2018-06-26 08:28:07 CDT; 1 day 5h ago
 Main PID: 4183 (code=exited, status=0/SUCCESS)
   CGroup: /system.slice/clmgr-power.service
           └─4942 clmgr-power /usr/bin/twistd --originalname -o -r poll --
logfile /opt/clmgr/log/clmgr-power.log --pidfile /var/ru...
.
.
.
```

If the daemon is not running, enter the following command to start the daemon:

```
# systemctl start clmgr-power
```

2. Use a text editor to open file `/opt/clmgr/log/clmgr-power.log`, which is the log file for the `clmgr-power` daemon on the admin node.

3. Verify that the log entries indicate a running daemon.

   For example, the following log file entries show that the `clmgr-power` daemon is running as expected:

```
2017-05-03 14:16:07+0000 [-] Log opened.
2017-05-03 14:16:07+0000 [-] twistd 14.0.2 (/usr/bin/python 2.7.9) starting up.
2017-05-03 14:16:07+0000 [-] reactor class: twisted.internet.pollreactor.PollReactor.
2017-05-03 14:16:07+0000 [-] Changing process name to clmgr-power
2017-05-03 14:16:07+0000 [-] Log opened.
2017-05-03 14:16:07+0000 [-] twistd 14.0.2 ( 2.7.9) starting up.
2017-05-03 14:16:07+0000 [-] reactor class: twisted.internet.pollreactor.PollReactor.
2017-05-03 14:16:07+0000 [-] PBServerFactory starting on 8800
.
.
.
```

   If the log file entries show traceback activity, the daemon might not be running correctly. If you see traceback entries, and you need help to interpret them, contact your technical support representative.

Verifying that the `smc-leaderd` daemon is running (HPE SGI 8600 clusters)

The following procedure explains how to make sure that the `smc-leaderd` daemon is running properly.

Procedure

1. Log into the admin node as the root user, and enter the following command to make sure that the `smc-leaderd` daemon is running:

```
# systemctl status smc-leaderd
```

   For example, the following example shows the daemon running as expected on a SLES system:

```
# systemctl status smc-leaderd
smc-leaderd.service - The SMC ADMIND Server
   Loaded: loaded (/usr/lib/systemd/system/smc-leaderd.service; enabled; vendor preset: enabled)
   Active: active (running) since Wed 2019-05-03 20:25:51 CDT; 20h ago
 Main PID: 15880 (twistd)
   CGroup: /system.slice/smc-leaderd.service
```

```
             15880 /usr/bin/python /usr/bin/twistd -o -r poll --logfile /var/log/smc-leaderd
             --pidfile /var/run/leaderd....

May 03 20:25:50 r1lead1 systemd[1]: Starting The SMC ADMIND Server...
May 03 20:25:51 r1lead1 twistd[15184]: /usr/lib64/python2.7/site-packages/twisted/internet/_sslverify.py:184: UserWarnin...
May 03 20:25:51 r1lead1 twistd[15184]: verifyHostname, VerificationError = _selectVerifyImplementation()
May 03 20:25:51 r1lead1 systemd[1]: Started The SMC ADMIND Server.
Hint: Some lines were ellipsized, use -l to show in full.
```

If the daemon is not running, start the daemon. Enter the following command:

```
# systemctl start smc-leaderd
```

2. Use a text editor to open file `/var/log/smc-leaderd`, which is the log file for the `smc-leaderd` daemon on the leader node.

3. Verify that the log entries indicate a running daemon.

   For example, the following log file entries show that the `smc-leaderd` daemon is running as expected:

```
2019-05-03 09:27:53-0500 [-] Log opened.
2019-05-03 09:27:53-0500 [-] twistd 14.0.2 (/usr/bin/python 2.7.9) starting up.
2019-05-03 09:27:53-0500 [-] reactor class: twisted.internet.pollreactor.PollReactor.
2019-05-03 09:27:53-0500 [-] QuietSite (TLS) starting on 911
2019-05-03 09:27:53-0500 [-] Starting factory 2019-05-03 09:27:53-0500 [-] PBServerFactory (TLS) starting on 901
2019-05-03 09:27:53-0500 [-] Starting factory 2019-05-03 09:27:53-0500 [-] F starting on '/tmp/admind.sock'
2019-05-03 09:27:53-0500 [-] Starting factory 2019-05-03 09:28:35-0500 [-] Network found! Updating....
2019-05-03 09:28:35-0500 [-].
.
.
.
```

   If the log file entries show traceback activity, the daemon might not be running correctly. If you see traceback entries, and you need help to interpret them, contact your technical support representative.

Verifying that the `icerackd` daemon is running (HPE SGI 8600 clusters)

The following procedure explains how to verify the `icerackd` daemon.

Procedure

1. From the admin node, use the `ssh` command to log into a leader node.

   For example, for `r1lead`, enter the following command:

```
# ssh r1lead
Last login: Wed Apr 15 18:43:16 2015 from admin
```

2. Enter the following command to verify that the `icerackd` daemon is running:

```
# systemctl status icerackd
```

   For example, the following output shows that the daemon is running as expected:

```
Checking for service icerackd                               done
```

   If the daemon is not running, start the daemon. Enter the following command:

```
# systemctl start icerackd
```

   Make sure to verify that the daemon is running on each leader node in the cluster.

3. Use a text editor to open file `/var/log/icerackd.log`, which is the log file for the power management daemon.

4. Use the `ps` command to verify that the log entries indicate a running daemon.

For example, the following output shows that the `icerackd` daemon is running as expected:

```
icerackd.service - LSB: Automatic blade discovery daemon
   Loaded: loaded (/etc/init.d/icerackd; bad; vendor preset: disabled)
   Active: active (running) since Wed 2019-04-25 11:12:53 CDT; 12min ago
     Docs: man:systemd-sysv-generator(8)
    Tasks: 2 (limit: 512)
   CGroup: /system.slice/icerackd.service
           └─5686 /usr/bin/python /usr/bin/twistd -l /var/log/icerackd.log
--pidfile /var/run/icerackd.pid -y /opt/sgi/lib/icerackd

Apr 25 11:12:52 r1lead systemd[1]: Starting LSB: Automatic blade discovery
daemon...
Apr 25 11:12:53 r1lead icerackd[5344]: Starting icerackd ..done
```

Make sure to repeat the preceding steps on each leader node in the cluster.

Using the `switchconfig` command

The `switchconfig` command displays switch settings and enables you to configure switches.

To retrieve help output online, which includes examples, enter the following:

```
# switchconfig --help
```

The preceding command displays all the possible subcommands. To retrieve more information about an individual subcommand, specify the following:

```
switchconfig subcommand --help
```

Cannot ssh from admin node to ICE compute nodes (HPE SGI 8600 clusters)

This problem arises when the admin node does not have a connection to the data network.

For more information, see the following:

Adjusting the domain name service (DNS) search order

Procedure

1. Log into the admin node as the root user.
2. Use the `cm node set` command in the following format to set the search path for the admin node:

   ```
   cm node set -n admin --domain-search-path subdomains
   ```

   For *subdomains*, specify one or more cluster subdomains, and separate each subdomain with a comma.

   For example:

   ```
   # cm node set -n admin --domain-search-path gbe.cm.lab.net,cm.lab.net,lab.net
   ```
3. Enter the `cm node set` command in the following format to set the search path for one of the leader nodes:

   ```
   cm node set -n leader_node --domain-search-path subdomains
   ```

   The variables are as follows:

- For *leader_node*, specify the hostname for one of the leader nodes.
- For *subdomains*, specify one or more cluster subdomains, and separate each subdomain with a comma.

For example:

```
# cm node set -n r1lead --domain-search-path head.cm.lab.net,ib0.cm.lab.net,lab.net
```

Repeat this step for each leader node in the cluster.

Compute nodes under a leader node are taking too long to boot

Symptom

Compute nodes under a leader node are taking too long to boot. This problem can exist on clusters with scalable unit (SU) leader nodes, on clusters with ICE leader nodes, or on compute nodes attached directly to the admin node as service nodes.

Cause

Generally, the 802.3ad (LACP) bonding mode provides more bandwidth and redundancy than the active-backup bonding mode. However, the bonding mode on a node must match the management Ethernet switch to which it is connected. The different leader node situations are as follows:

- For ICE leader nodes, this bonding mode defaults to 802.3ad (LACP).
- For scalable unit (SU) leader nodes, configure the bonding mode in the cluster configuration file. For example, the configuration file lines could look like this:

```
internal_name=service1, hostname1=leader1, mgmt_net_interfaces=”eno1,eno2”, mgmt_net_bonding_master=bond0,
mgmt_net_bonding_mode=802.3ad, predictable_net_names=yes, mgmt_net_macs=”aa:bb:cc:dd:ee:11,aa:bb:cc:dd:ee:12”
.
.
.
```

If the following are all true, use the procedure in this topic to verify the bonding mode and if necessary, to update the bonding mode:

- The discover command has run
- The node in question is configured into the cluster
- You need to change the bonding mode for the node

The procedure works on any type of node (admin node, leader node, non-ICE compute node, or ICE compute node).

Action

1. Log into the admin node as the root user.
2. Use the `cadmin` command in the following format to display the bonding mode:

   ```
   cadmin --show-mgmt-bonding --node hostname
   ```

   For *hostname*, specify the hostname of the node you want to verify.

   Example 1:

   ```
   admin~# cadmin --show-mgmt-bonding --node leader1
   active-backup
   ```

Example 2:

```
admin~# cadmin --show-mgmt-bonding --node r1lead
802.3ad
```

3. Use the `cm node set` command to reset the bonding mode on a given node.

   Use the command in one of the following formats:

   ```
   cm node set -n hostname --mgmt-bonding active-backup
   ```

   Or

   ```
   cm node set -n hostname --mgmt-bonding 802.3ad
   ```

   Example 1:

   ```
   admin# cm node set -n n0 --mgmt-bonding 802.3ad
   ```

   Example 2:

   ```
   admin~# cm node set -n leader1 --mgmt-bonding 802.3ad
   ```

4. Reset the leader node.

   Use the cm power reset command in the following format:

   ```
   cm power reset -t target_type hostname
   ```

   For *target_type*, specify one of the following:

target_type	**Appropriate node types**
`node`	Non-ICE compute nodes, ICE compute nodes
`leader`	Scalable unit (SU) leader nodes, ICE leader nodes

   Example 1:

   ```
   admin~# cm power reset -t node n0
   ```

   Example 2:

   ```
   admin~# cm power reset -t leader leader1
   ```

   Wait for the node to reboot fully.

5. Use the `switchconfig_configure_node` command in the following format to configure the management switch attached to the node:

   ```
   switchconfig_configure_node --node hostname
   ```

   Example 1:

   ```
   # switchconfig_configure_node --node leader1
   ```

   Example 2:

   ```
   # switchconfig_configure_node --node r1lead
   ```

Cannot find the management switch that a node is plugged into

Symptom

Cannot find the management switch that a node is plugged into

Action

1. From the admin node, use the `arp` command to find the MAC address of the node.

   The command format is as follows:

   ```
   arp hostname
   ```

   For *hostname*, enter the hostname of the node.

   For example:

   ```
   admin:~ # arp r1lead
   Address                  HWtype          HWaddress           Flags Mask            Iface
   r1lead                   ether           00:25:90:96:4e:ac   C                     bond0
   ```

2. Use the `switchconfig find` command.

   The `switchconfig find` command returns the switch upon which a given node MAC address exists. The command searches multiple switches and displays information about physical ports and switches.

   NOTE: Some long commands in this topic use the \ character to continue the command to a second line.

   Example 1. The following command searches all management switches for MAC address `00:25:90:96:4e:ac`:

   ```
   admin:~ # switchconfig find --switches all --macs 00:25:90:96:4e:ac
   mac-address              switch        find_method     port
   -------------------------------------------------------------------
   00:25:90:96:4e:ac        mgmtsw3        lldp           1:1
   ```

   The preceding output shows the following:

   - The MAC address is found on `mgmtsw3`, port `1:1`.
   - The command used the link layer discovery protocol (LLDP) to determine its findings.

ICE leader nodes cannot ping their chassis management controllers

Symptom

ICE leader nodes cannot ping their chassis management controllers.

This problem can occur in the following situations:

- The management switch configuration was reset to factory settings.
- A power outage occurred, and the management switch lost its configuration. This outcome occurs when the configuration was not saved.

The result is that the ports that are connected to the CMCs or the leader nodes are in one of the following states:

- The ports do not have the proper bonding
- The ports do not have the proper VLAN configuration

Proper bonding and proper VLAN configuration are required for proper communication.

Solution 1

Cause

The cmcdetectd service is not running, or the cmcdetectd service failed to configure the switches.

Action

1. Use less to open the following log file, and check the error messages or warning messages related to the cmcdetectd service:

   ```
   /var/log/cmcdetectd.log
   ```

2. Check the messages in /var/log/cmcdetectd.log.
3. (Conditional) Restart the cmcdetectd service.

   Complete this step if the cmcdetectd.log messages indicate that the cmcdetectd service failed to configure the management switches that the chassis management controllers (CMCs) plug into.

   Enter the following commands to restart the service and monitor the restart:

   ```
   # systemctl restart cmcdetectd.service
   # tail -f /var/log/cmcdetectd.log
   ```

4. Evaluate the results, and if problems persist, complete the steps in Solution 2.

Solution 2

Action

1. Open the following file and verify that the switch configuration information is correct:

   ```
   /etc/cmc-switch-info.txt
   ```

2. Enter the following commands to stop the service and to direct the service to use the information in the switch configuration information file:

   ```
   # systemctl stop cmcdetectd.service
   # cmcdetectd --switchconfig
   ```

3. Wait for the cmcdetectd command to complete.
4. Enter the following commands to restart the cmcdetectd service and monitor the restart:

   ```
   # systemctl start cmcdetectd.service
   # tail -f /var/log/cmcdetectd.log
   ```

Solution 3

Cause

The management switch that is connected to the leader node is not configured properly.

Action

1. Enter the following command:

   ```
   switchconfig_configure_node --node hostname
   ```

For example:

```
# switchconfig_configure_node --node r1lead
```

If necessary, use the following additional arguments:

- `--daemon`

  Runs the command in the background. View `/var/log/switchconfig.log` to monitor progress.

- `--dry-run`

  Does not run any `switchconfig` commands, but prints out commands that otherwise would have been run.

- `--debug`

  Logs verbose output to the following file:

  `/var/log/switchconfig.log`

Restarting the `blademond` daemon (HPE SGI 8600 clusters)

Procedure

1. From the admin node, use the `ssh` command to log into one of the leader nodes.
2. Enter the following command to stop the daemon:

   ```
   r1lead1:~ # crm resource stop p_blademond
   ```

3. Enter the following command to remove `/etc/dhcpd.conf.d/ice.conf` or `/etc/dhcp/dhcpd.conf.d/ice.conf`:

   ```
   r1lead1:~ # rm ice.conf dhcpd.conf
   ```

4. Enter the following command to remove `slot_map`:

   ```
   r1lead1:~ # rm /var/opt/sgi/lib/blademond/slot_map
   ```

5. Enter the following command to start the daemon:

   ```
   r1lead1:~ # crm resource start p_blademond
   ```

Log files

All of the log files reside in the `/var/log` directory. The main log files are as follows:

- `/var/log/messages`
- On leader nodes, `/var/log/dhcpd` (clusters with leader nodes only)

The following are some other log files in the `/var/log` directory that might interest you:

- `/var/log/blademond` (HPE SGI 8600 clusters)

On the leader nodes, this file shows the blademond daemon actions, including blade changes, calls to discover-rack, and so on. If there are chassis management controller (CMC) communication problems, they often appear in this log.

- /var/log/cmcdetectd.log

  On the admin node, cmcdetectd logs its actions as it configures the switches for CMCs in the system. Watch for progress or errors here.

- /var/log/dhcpd

  This file contains DHCP messages.

  On clusters with leader nodes, a file by this name resides on both the admin node and on each leader node.

  On clusters without leader nodes, this file resides on the admin node.

- /var/log/discover-rack (HPE SGI 8600 clusters)

  On the admin node, the discover-rack call is facilitated by blademond when new nodes are found. Problems with node configuration often appear in this log file.

- /var/log/switchconfig.log

  On the admin node, there is a switchconfig command-line tool. This tool is largely used by the discover command as nodes are configured. Its actions are logged in this log file. If leader node VLANs are not functioning properly, check the switchconfig log file.

- Log files related to the scalable unit (SU) leader node infrastructure reside in the following directories:

  - ctdb log files reside in /var/log/log.ctdb
  - Gluster log files reside in /var/log/glusterfs/*
  - Gluster bricks reside in /var/log/glusterfs/bricks/*

Chassis management controller (CMC) slot_map / blademond debugging hints (HPE SGI 8600 clusters)

This topic describes what to do when the blademond daemon cannot find a system blade.

To explore the problem, first issue a ping command to the CRCs, and then proceed to one of the following:

- If the CMCs do not answer the ping command, see the following:

  ICE leader nodes cannot ping their chassis management controllers

- If the CMCs answer the ping command, verify that they have a valid slot map.

  If the slot map returned by the CMC is missing entries, then blademond daemon cannot function properly. The daemon operates on information passed to it by the CMC. The following are some commands you can run from the leader node:

  - To dump the slot map from each CMC to your screen, enter the following command:

    ```
    r1lead:~ # /opt/sgi/lib/dump-cmc-slot-tables
    ```

  - To query an individual slot map, enter the following command:

    ```
    r1lead:~ # echo STATUS | netcat r1i0c 4502
    ```

NOTE: In some software distributions, `netcat` is `nc`.

If the CMCs have valid slot maps, then you can focus on how `blademond` is functioning. To turn on debug mode in the `blademond` daemon, use the following command to send the daemon a `SIGUSR1` signal from the leader node, as follows:

```
# kill -USR1 pid
```

To disable debug mode, send it another `SIGUSR1` signal. In the `blademond` log file, look for messages that describe debug mode being enabled or disabled. The `blademond` daemon maintains the slot map in `/var/opt/sgi/lib/blademond/slot_map` on the leader nodes. The slot map appears in `/var/opt/sgi/lib/blademond/slot_map.`*`rack_number`* on the admin node.

For a `blademond --help` statement, `ssh` into the `r1lead` leader node, as follows:

```
[root@admin ~]# ssh r1lead
Last login: Tue Jan 17 13:21:34 2012 from admin
[root@r1lead ~]#
[root@r1lead ~]# /opt/sgi/lib/blademond --help
Usage: blademond [OPTION] ...

Discover CMCs and blades managed by CMCs.

Note: This daemon normally takes no arguments.
  --help      Print this usage and exit.
  --debug     Enable debug mode (also can be enabled by setting CM_DEBUG)
  --fakecmc   Development only: Discover fake CMCs instead of real ones
  --scan-once Initialize, scan for blades, set blades up. Do not daemonize.
              Do not keep looping - do one pass and exit.
```

Resolving chassis management controller (CMC) slot map ordering problems (HPE SGI 8600 clusters)

If either of the following conditions exist, the CMC slot map might be corrupted:

- There are `ssh` key failures.

  or

- The compute node hosts seem to be BMCs.

The CMC maintains a cache file. The file records the MACs that are BMC MACs and the MACs that are host MACs. The CMC uses this information, combined with switch port location information in the embedded Broadcom switch, to generate the slot map used by the `blademond` daemon.

Certain situations, for example, a CMC reflash, might remove the cache file but leave CMC power active. In this situation, the CMC cannot determine which MACs on a given embedded switch port are MAC hosts and which are BMC hosts. Because of the confusion, the CMC gets the order randomly incorrect. It then caches the incorrect order.

The following procedure explains how to fix this problem. For each CMC, you use `pfctl` to power off the CMC, zero out the MAC cache file, reset the CMC, and then restart `blademond`.

Procedure

1. `ssh` as root to the leader node, as follows:

   ```
   sys-admin:~ # ssh r1lead
   Last login: Thu Jan 26 13:57:53 2012 from admin
   r1lead:~ #
   ```

2. Enter the following command to disable the `blademond` daemon:

   ```
   r1lead:~ # systemctl stop blademond
   ```

3. Turn off chassis power for each CMC using the `PDSH` command, as follows:

   ```
   # PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc pfctl off
   ```

4. Zero out the slot map cache file, as follows:

   ```
   # PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc cp /dev/null /work/net/broadcom_mac_addr_cache
   ```

5. Reboot the CMC, as follows:

   ```
   # PDSH_SSH_ARGS_APPEND="-F /root/.ssh/cmc_config" pdsh -g cmc reboot
   ```

6. Restart `blademond` from scratch.

   For information, see the following:

   Restarting the `blademond` daemon (HPE SGI 8600 clusters)

Ensuring that the hardware clock has the correct time

Some software distributions do not synchronize the system time to the hardware clock as expected. As a result, the hardware clock is not synchronized with the system time, which is the correct condition. At shutdown, the system time is copied to the hardware clock, but sometimes this synchronization does not happen.

To set the compute node hardware clocks properly, check the following:

- Make sure that the admin node and the leader nodes have the correct time.
- Use the `chronyc sources` command to show synchronization. For example:

  ```
  [root@admin ~]# chronyc sources
  210 Number of sources = 2
  MS Name/IP address         Stratum Poll Reach LastRx Last sample
  ===============================================================================
  ^* admin.head.cm.eag.rdlabs>     8   5    77     6   +155ns[+2243ns] +/- 3649ns
  ^? toddadev.your.org             0   5     0     -     +0ns[   +0ns] +/-    0ns

  [root@admin ~]# ssh r1lead chronyc sources
  210 Number of sources = 1
  MS Name/IP address         Stratum Poll Reach LastRx Last sample
  ===============================================================================
  ^* admin                         9   5   377    30  +3204ns[+4416ns] +/-   51ms

  [root@admin ~]# ssh n0 chronyc sources
  210 Number of sources = 1
  MS Name/IP address         Stratum Poll Reach LastRx Last sample
  ===============================================================================
  ^* admin.head.qe-cm.chf.rdl>     9   6   377    28  -3810ns[-4516ns] +/-  152ms
  ```

  In the preceding output, note the following:

  - The carat (^) adjacent to the hostname or IP address shows that the node is an NTP server.
  - The asterisk (*) shows the NTP server to which the system is synchronized.
  - The plus sign (+) shows an NTP server that is a combined source.
  - The minus sign (–) shows an NTP server that is not a combined source.

- Use the `chronyc tracking` command to show the state of a node. For example:

```
[root@admin ~]# chronyc tracking
Reference ID    : AC170001 (admin.head.cm.eag.rdlabs.hpecorp.net)
Stratum         : 9
Ref time (UTC)  : Thu Nov 07 23:34:55 2019
System time     : 0.000000000 seconds slow of NTP time
Last offset     : +0.000002088 seconds
RMS offset      : 0.000002088 seconds
Frequency       : 0.129 ppm slow
Residual freq   : +0.028 ppm
Skew            : 0.000 ppm
Root delay      : 0.000006853 seconds
Root dispersion : 0.000331546 seconds
Update interval : 0.0 seconds
Leap status     : Normal

[root@admin ~]# ssh r1i0n0.gbe chronyc sources
210 Number of sources = 1
MS Name/IP address         Stratum Poll Reach LastRx Last sample
===============================================================================
^* r1lead.gbe.cm.eag.rdlabs>    10  10   377   879  -4261ns[ +264ns] +/-   641us
```

- To set the hardware clock to the system clock, enter the following command:

```
admin:~ # hwclock --systohc
```

- To set the hardware clock to the system clock on the leader nodes and on the non-ICE compute nodes, enter the following commands:

```
admin:~ # pdsh -g leader hwclock --systohc
admin:~ # pdsh -g compute hwclock --systohc
```

- To confirm the current hardware clock time, enter the `hwclock` command without options, as follows:

```
sys-admin:~ # hwclock
Thu 26 Jan 20XX 10:57:27 PM CST  -0.750431 seconds
```

- To confirm the current hardware clock on leader nodes and non-ICE compute nodes, enter the following commands:

```
admin:~ # pdsh -g leader date
r1lead: Tue Apr 18 15:00:20 PDT 20XX
admin:~ # pdsh -g compute date
node0: Tue Apr 18 15:00:45 PDT 20XX
```

Troubleshooting a leader node with misconfigured switch information (HPE SGI 8600 clusters)

Typically, as you discover leader nodes, the installer calls `switchconfig` automatically, and the switch ports associated with the leader node are configured as follows:

- Default VLAN 1
- Accept rack VLAN packets tagged (rack 1 `vlan` is `vlan101`)
- Link aggregation is the bonding mode between the two ports associated with the leader node

If the following situations arise, you can run the `switchconfig` command by hand to configure the switch:

- If you move a leader node
- If `switchconfig` fails during discovery

The following procedure explains how to configure a leader node switch.

1. Review the switch configuration rules.

   For information, see the following:

   Switch wiring rules (HPE SGI 8600 clusters)

2. Make sure that the admin node can reach all the management switches.
3. Find the MAC addresses associated with the leader node NICs.

   For example, run the following command on the leader node in question:

   ```
   r1lead:~ # cat /proc/net/bonding/bond0
   Ethernet Channel Bonding Driver: v3.7.1 (April 27, 2019)

   Bonding Mode: fault-tolerance (active-backup)
   Primary Slave: None
   Currently Active Slave: en01
   MII Status: up
   MII Polling Interval (ms): 100
   Up Delay (ms): 0
   Down Delay (ms): 0

   Slave Interface: en01
   MII Status: up
   Speed: 1000 Mbps
   Duplex: full
   Link Failure Count: 0
   Permanent HW addr: 00:25:90:02:49:5a
   Slave queue ID: 0

   Slave Interface: en02
   MII Status: up
   Speed: 1000 Mbps
   Duplex: full
   Link Failure Count: 0
   Permanent HW addr: 00:25:90:02:49:5b
   Slave queue ID: 0
   ```

CAUTION: Bonded NICs are in play, so you cannot use the `ip` command to retrieve both MAC addresses. The `ip` command shows the same MAC address for both Ethernet networks. Use the `cat` command that this step shows to obtain the MAC addresses.

4. Enter the following command to determine the management switches that are present:

```
r1lead:~ # cnodes -mgmtsw
mgmtsw0
```

5. When you have the list of management switches and the MAC addresses of the leader nodes, use the `switchconfig` command to set the switch.

   For example:

```
# switchconfig set -s mgmtsw0 -d 1 -v 101 -b lacp -m 00:0e:ed:0a:f2:0d,00:0e:ed:0a:f2:0e -r yes
```

   The preceding command replaces the MACs and management switches with the proper ones. It replaces the `101` with the VLAN for the rack, which is typically 100 + the rack number. For example, rack 1 is 101, and rack 2 is 102.

Switch wiring rules (HPE SGI 8600 clusters)

Some clusters have a redundant management network (stacked pairs of switches). Other clusters have cascaded switches, in which switch stacks are cascaded from the top-level switch. When discovering cascaded switches, it is impossible to know the connected switch ports of all trunks in advance. When using the `discover` command to configure cascaded switches, you start with only one cable and add the second one later on.

When trunks are configured, it is often hard to find the MAC address of both legs of the trunk. The difficulty arises because the trunked connection just uses one MAC for the connection. Therefore, you can rely on rules that infer the second port connection based on the first port connection.

The following are some simple wiring rules:

- In a redundant management network (RMN) configuration, use the same port number in both switches for a particular piece of equipment. That is, make sure to assign the same port number in each stack to the following components:

  - Admin nodes
  - Leader nodes
  - Non-ICE compute nodes with services installed upon them
  - CMCs

  In other words, if you connect `r1lead`'s first NIC to switch A, port 43, then you must connect `r1lead`'s second NIC to switch B, port 43. Likewise, if you connect CMC `r1i0c` CMC-0 port to switch A, port 2, then `r1i0c` CMC-1 port must go to switch B port 2.

- When adding cascaded switch stacks, all switch stacks must cascade from the primary switch stack. In other words, there is always only, at most, one switch hop.

- When discovering cascaded switches pairs in an RMN setup, observe the following:

- If you are connecting switch stack 1, switch A, port 48 to switch stack 2, then connect the second trunked connection to stack 2, switch B, port 48.
- Until the cascaded switch stack is discovered, you must leave one trunk leg unplugged temporarily to prevent looping.
- The `discover` command tells you when it is safe to plug in the second leg of the trunk. This notification avoids circuit loops.

Bringing up the second NIC in an admin node when it is down

The logical interface, `bond0`, can contain one or more physical NICs. It is possible for these physical NICs to be administratively down or unplugged. The following procedure explains how to determine link status of the physical NICs under bond0.

The following procedure explains how to detect this situation.

Procedure

1. Check the Ethernet port of the add-in card and confirm that it is lit.
2. Confirm that the add-in card connection to the management switches is using port `0`.

   Make sure that port `1` is not connected.

   This step verifies the wiring.
3. Examine the following file to see whether the second, redundant Ethernet interface link is down:

   ```
   /proc/net/bonding/bond0
   ```
4. Use the `ethtool` command to determine if the content of the `Link detected:` field is `no`.

   For example:

   ```
   # ethtool management_interface1
   ```
5. Enter the following command to bring up the interface:

   ```
   # ip link set management_interface1 up
   ```
6. To verify that the link is detected, run the following command:

   ```
   # ethtool management_interface2
   ```
7. In the preceding command output, search for `yes` in the `Link detected` field.

Booting nodes with iPXE

If a node fails to boot, specify that iPXE load first and that iPXE load GRUB version 2.

From the admin node, enter the following command to specify the nodes:

```
cadmin --set-dhcp-bootfile --node node_ID ipxe
```

For *node_ID*, specify the identifier of the node that did not boot. For example, for a compute node, specify its hostname.

To verify whether a node is enabled to load iPXE first, enter the following command:

```
cadmin --show-dhcp-bootfile
```

The DHCP log file messages reside in the following file on the admin node:

```
/var/log/dhcp
```

Miniroot operations

The following topics can help you troubleshoot a suspected miniroot kernel problem:

- **About the miniroot**
- **Entering rescue mode**
- **Logging into the miniroot to troubleshoot an installation**

About the miniroot

The cluster manager miniroot is a small Linux environment based on the same RPM repositories that generated the root image itself. The cluster manager software uses the miniroot to install the software and to boot the following nodes over the cluster network:

- Leader nodes and the nodes under their control
- Non-ICE compute nodes with services installed upon them

The miniroot is a small, bootable file system. It includes kernel modules such as disk drivers, Ethernet drivers, and other software. These software drivers are associated with a specific kernel number. As new driver updates become available, the operating systems distribute additional kernels. The system requires at least one kernel to be associated with a specific node image. You can associate more than one kernel with a specific node image.

When the cluster manager boots a node, the cluster manager uses the images that reside in the admin node image repository. Because the nodes boot over the network, it is important that the images in the admin node image repository include the correct kernels. That is, it is important that the following are identical:

- The kernel in the on-node image. This image is the installed image that resides on the node while the node is running.
- The kernel in the node image repository on the admin node. There can be multiple node images for a single node type in the image repository.
- The kernels in the kernel repository on the admin node. There can be multiple kernels in the kernel repository. The `cm node provision` command includes a kernel from the repository when it builds a node image. These kernels reside in the following directory on the admin node:

  ```
  /opt/clmgr/tftpboot/images
  ```

Use the cluster manager `cm node provision` command to update images. When you use this command, the cluster manager ensures synchronization between the on-node image and the image in the admin node image repository. Do not change an on-node image manually without using the cluster manager commands. If you omit the command and subsequently boot the node, one of the following occurs:

- The boot fails

  or

- The cluster manager detects a mismatch between the following:
  - The kernel loaded over the network
  - The kernel and associated modules in the image itself

The mismatch can lead to a node that boots but has no network, for example. Therefore, it is important that all the images in the image repository on the admin node contain the on-node images with all the kernels in use.

If you update any images manually, use the following command:

```
cm image update -i image -k
```

The preceding command has the following effects:

- The command synchronizes the kernels and the `initrd` daemon in the images.
- The command writes a copy of the kernel to the `/opt/clmgr/tftpboot/images` directory for future use when performing network boots.

Entering rescue mode

To go into miniroot rescue mode, enter commands such as the following:

```
# cm node set -n n1 --kernel-extra-params "rescue=1"
# cm node refresh netboot -n n1
```

The preceding command includes the `rescue=1` kernel command-line argument. This argument ensures that the kernel command line includes `rescue=1`.

To remove `rescue=1`, use commands such as the following:

```
# cm node unset --kernel-extra-params -n n1
# cm node refresh netboot -n n1
```

Logging into the miniroot to troubleshoot an installation

The miniroot brings up an `ssh` server for its operations. If an installation fails, first look to the serial console using the `conserver` command and any console log files.

To examine the situation from a separate session, specify port 40. The miniroot environment listens for `ssh` connections on port 40.

For example, assume that the node that failed to install is `r01n02`. The following command logs you into the miniroot on node `r01n02` from the admin node:

```
admin# ssh -p 40 root@r01n02
miniroot#
```

At this point, you can run typical Linux commands to debug the problem. HPE supports only a subset of the standard Linux commands on the miniroot.

To connect to an ICE compute node in a similar way, use the `ssh` command as follows:

1. Log into the leader node that manages the ICE compute node.
2. Use the `ssh` command again to log into port 40 on the specific ICE compute node.

Troubleshooting an HA admin node configuration

The following list shows the commands that you can use to troubleshoot an HA admin node configuration problem:

- To verify the network configuration, examine the `/etc/hosts` file.

For example:

```
# cat /etc/hosts
137.38.97.22    acme-admin1
137.38.97.31    acme-admin2
192.168.0.1     acme-admin1-ptp
192.168.0.2     acme-admin2-ptp
172.23.254.253  acme-admin1-head
172.23.254.254  acme-admin2-head
137.38.97.109   acme-admin1-bmc
137.38.97.104   acme-admin2-bmc
```

- To verify the firewall, use the following commands:
  - On RHEL platforms, enter the following command:

    ```
    # cat /etc/firewalld/zones/public.xml | grep service
    <service name="dhcpv6-client"/>
    <service name="ssh"/>
    <service name="high-availability">
    ```

  - On SLES platforms, enter the following command:

    ```
    # cat /etc/sysconfig/SuSEfirewall2 | grep FW_CONFIGURATIONS_EX
    FW_CONFIGURATIONS_EXT="cluster sshd vnc-server"
    ```

Troubleshooting UDPcast transport failures from the admin node or from a leader node

You might encounter one of the following situations when you use the UDPcast (multicast) transport method during installation:

- The client side waits forever for a `udp-receiver` process to complete.
- The `udp-receiver` processes repeatedly attempts to provision a node.

If either of the preceding conditions exist, you have an error situation.

The `systemimager-server-flamethrowerd` service manages the `udp-sender` instances. The following procedure explains how to remedy this situation by stopping and restarting UDPcast `flamethrower` services.

The following procedure explains another UDPcast troubleshooting strategy:

Troubleshooting UDPcast transport failures from the switch

Procedure

1. As the root user, log in to the node that serves UDPcast.

   For example, log into the admin node if your goal is to restart UDPcast services for leader nodes or non-ICE compute nodes. Likewise, log into the ICE leader node if your goal is to restart UDPcast services for ICE compute nodes.

2. Stop the `systemimager-server-flamethrowerd` services.

   Proceed as follows:

- From an admin node, a highly available (HA) admin node, or a non-HA leader node, enter the following command:

  ```
  # systemctl stop systemimager-server-flamethrowerd
  ```

- From an HA leader node, enter the following command:

  ```
  # crm resource stop p_flamethrowerd
  ```

3. Enter the following command to check for `udp-sender` processes that did not stop:

   ```
   # ps -ef | grep udp-sender
   ```

4. (Conditional) Enter one or more `kill -9 process_ID` commands to stop `udp-sender` processes that are still running.

5. Start the `systemimager-server-flamethrowerd` service.

   Type one of the following commands:

   - From an admin node, an HA admin node, or a non-HA leader node, enter the following command:

     ```
     # systemctl start systemimager-server-flamethrowerd
     ```

   - From an HA leader node, enter the following command:

     ```
     # crm resource start p_flamethrowerd
     ```

Troubleshooting UDPcast transport failures from the switch

UDPcast relies on IGMP technology. The IGMP technology determines the physical ports that subscribe to specific multicast addresses at layer 2 (data link) in the OSI model. In some scenarios, IGMP can be problematic for UDPcast.

The following `switchconfig` commands show the parameters that retrieve IGMP status information:

- To view global IGMP status on a management switch:

  ```
  switchconfig igmp --switches mgmtswX --info
  ```

- To the IGMP status for a specific VLAN on a management switch:

  ```
  switchconfig igmp --switches mgmtswX --info --vlan VLAN_#
  ```

You can disable IGMP on the management switches. Disabling and enabling IGMP have the following effects:

- When IGMP is enabled, a layer-2 multicast tree is created on the Ethernet switches. This tree determines the ports to which the UDPcast traffic is forwarded. In some cases, in the UDPcast code, the `IGMP Join` packets from the `udp-receiver` clients do not reach the Ethernet switches. In these cases, the multicast tree is not formed.
- When IGMP is disabled globally, the Ethernet switches convert all multicast packets to broadcast packets. In this case, the packets are nearly guaranteed to reach every host in a VLAN. Thus, the reduced performance increases reliability.

The following `switchconfig` commands show the parameters that disable IGMP on the management switches:

- To disable IGMP globally on a management switch:

  ```
  switchconfig igmp --switches mgmtswX --disable
  ```

- To disable IGMP on a specific VLAN on a management switch:

  ```
  switchconfig igmp --switches mgmtswX --disable --vlan VLAN_#
  ```

To re-enable IGMP on global or per-VLAN basis, replace `--disable` with `--enable`. In addition, if necessary, use the `--version` parameter to specify the IGMP version. You can specify `--version 2` or `--version 3` . The default version is version 2.

The following command re-enables IGMP with IGMP version 3 on `mgmtsw0`:

```
# switchconfig igmp --switches mgmtsw0 --enable --version 3
```

Reprovisioning scalable unit (SU) leader nodes

Procedure

1. Make sure that all SU leader nodes are set to boot over the network.

   Complete the following steps:

   a. Use the `ssh` command to log into one of the SU leader nodes.

   b. Use the `efibootmgr` command to list the boot order for the devices on the node.

      For example:

      ```
      # efibootmgr
      BootCurrent: 000E
      Timeout: 0 seconds
      BootOrder: 000E,
      0018,0000,0001,0002,0003,0004,0005,0006,0007,0008,0009,000A,000B,
      0014,0015,0016,0017,000C,0011,0013,0012,0020,000F,001B,0010,001F,001A,
      001E,000D,0021,0022
      Boot0000* System Utilities
      Boot0001 Embedded UEFI Shell
      Boot0002 Diagnose Error
      Boot0003 Intelligent Provisioning
      Boot0004 Boot Menu
      Boot0005 Network Boot
      Boot0006 View Integrated Management Log
      Boot0007 HTTP Boot
      Boot0008 PXE Boot
      Boot0009 Embedded Diagnostics
      Boot000A* Generic USB Boot
      Boot000B* Internal SD Card 1 : Generic USB3.0-CRW
      Boot000C* Embedded RAID 1 : HPE Smart Array P408i-a SR Gen10 - 2.1 TiB,
      RAID1+0 Logical Drive 1(Target:0, Lun:0)
      Boot000D* Slot 10
      Boot000E* SGI Slot Chooser
      Boot000F* Slot 3
      Boot0010* Slot 4
      ```

```
Boot0011* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port
841QSFP28 Adapter - HCA (HTTP(S) IPv4)
Boot0012* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port
841QSFP28 Adapter - HCA (HTTP(S) IPv6)
Boot0013* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port
841QSFP28 Adapter - HCA (PXE IPv4)
Boot0014* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter -
NIC (HTTP(S) IPv4)
Boot0015* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter -
NIC (HTTP(S) IPv6)
Boot0016* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter -
NIC (PXE IPv6)
Boot0017 Intelligent Provisioning
Boot0018* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter -
NIC (PXE IPv4)
Boot0019* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter -
NIC (PXE IPv6)
Boot001A* Slot 1
Boot001B* Slot 8
Boot001C* Windows Boot Manager
Boot001D* Windows Boot Manager
Boot001E* Slot 2
Boot001F* Slot 6
Boot0020* Slot 7
Boot0021* Slot 5
Boot0022* Slot 9
```

c. Use the `efibootmgr` command to specify that the device identified in the previous step is first in the boot order list.

For example:

```
# efibootmgr -o 0018,000E,0000,0001,0002,0003,0004,0005,0006,0007,0008,0009,000A,000B,0014,0015,0016,\
0017,000C,0011,0013,0012,0020,000F,001B,0010,001F,001A,001E,000D,0021,0022
```

d. Enter the `efibootmgr` command to verify that the correct boot order is set.

For example:

```
# efibootmgr
BootCurrent: 000E
Timeout: 0 seconds
BootOrder: 0018,000E,0000,0001,0002,0003,0004,0005,0006,0007,0008,0009,000A,000B,0014,0015,0016,0017,000C,0011,
0013,0012,0020,000F,001B,0010,001F,001A,001E,000D,0021,0022
Boot0000* System Utilities
Boot0001 Embedded UEFI Shell
Boot0002 Diagnose Error
Boot0003 Intelligent Provisioning
Boot0004 Boot Menu
Boot0005 Network Boot
Boot0006 View Integrated Management Log
Boot0007 HTTP Boot
Boot0008 PXE Boot
Boot0009 Embedded Diagnostics
Boot000A* Generic USB Boot
Boot000B* Internal SD Card 1 : Generic USB3.0-CRW
Boot000C* Embedded RAID 1 : HPE Smart Array P408i-a SR Gen10 - 2.1 TiB, RAID1+0 Logical Drive 1(Target:0, Lun:0)
Boot000D* Slot 10
Boot000E* SGI Slot Chooser
Boot000F* Slot 3
Boot0010* Slot 4
Boot0011* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port 841QSFP28 Adapter - HCA (HTTP(S) IPv4)
Boot0012* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port 841QSFP28 Adapter - HCA (HTTP(S) IPv6)
Boot0013* Slot 1 Port 1 : HPE InfiniBand EDR/Ethernet 100Gb 2-port 841QSFP28 Adapter - HCA (PXE IPv4)
Boot0014* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter - NIC (HTTP(S) IPv4)
Boot0015* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter - NIC (HTTP(S) IPv6)
```

```
Boot0016* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv6)
Boot0017 Intelligent Provisioning
Boot0018* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv4)
Boot0019* Embedded LOM 1 Port 1 : HPE Ethernet 1Gb 4-port 331i Adapter - NIC (PXE IPv6)
Boot001A* Slot 1
Boot001B* Slot 8
Boot001C* Windows Boot Manager
Boot001D* Windows Boot Manager
Boot001E* Slot 2
Boot001F* Slot 6
Boot0020* Slot 7
Boot0021* Slot 5
Boot0022* Slot 9
```

2. Enter the following command to power off the cluster:

   ```
   # cm power off -t system
   ```

3. Use the following command to provision the SU leader nodes with the new image:

   ```
   cm node provision -n su_leader_hostnames
   ```

   For *su_leader_hostnames*, specify the names of the SU leader nodes.

   For example:

   ```
   # cm node provision -n "leader*"
   ```

4. Enter the following command to verify that the SU leader nodes are booted:

   ```
   # cm power status -t system
   ```

5. Run the `su-leader-setup`, the `enable-su-leader`, and `activate-nfs-image` commands as shown in the following procedure:

   Configuring the Gluster file system and completing the cluster configuration

   The final steps in the preceding procedure run the `discover` command and back up the cluster configuration. You do not have to complete those steps.

6. Reprovision the non-ICE compute nodes:

   ```
   # cm node provision -n "node*"
   ```

Troubleshooting the `clmgr-power` service and the `cmcinventory` service (HPE Apollo 9000)

You can use the `rest_agent_tool` command to help you troubleshoot the following services:

- `clmgr-power`
- `cmcinventory`

Enter the following command to retrieve information about `rest_agent_tool`.

```
# rest_agent_tool -h
```

NOTE: Use this command only while working with an HPE representative.

Troubleshooting the `cmcinventory` service (HPE Apollo 9000)

You can examine the contents of the inventory files that the cluster manager uses when it automatically configures non-ICE compute nodes into the cluster. The inventory files reside in the following directory:

```
/opt/clmgr/cmcinventory/inventory
```

Enter the following command to see the names of the inventory files in the directory:

```
# ls /opt/clmgr/cmcinventory/inventory/
 inventory.r1c1 inventory.r1c2 inventory.r2c1 inventory.r2c2
```

You can examine the content of the inventory files in the directory and compare them with what is in the `rest_agent_tool`.

If `cmcinventory` is not running, use the `rest_agent_tool` command to obtain a current copy. The command format is as follows:

```
rest_agent_tool -b cmc_ID -I
```

For *cmc_ID*, specify the chassis manager controller (CMC) IP address or the CMC hostname.

For example:

```
# rest_agent_tool -b r1c1 -I
```

The command generates the following file:

```
/opt/clmgr/cmcinventory/conf/fastdiscover.conf
```

There are other files in the `/opt/clmgr/cmcinventory` directory, such as `flashnodes.conf`. The `flashnodes.conf` file helps bring up and configure nodes.

Bad firmware can cause some nodes to have incomplete MAC address information. The `cmcinventory` service attempts to generate a valid MAC address, and it stores the nodes with bad information to `flashnodes.conf`. Some `flashnodes.conf` files include a time stamp. Files with a time stamp are older and can be deleted.

Connecting to the virtual admin node in a cluster with a highly available (HA) admin node

Procedure

1. Log into one of the physical admin nodes as the root user.
2. Enter the following command in a terminal window on the physical admin node:

   ```
   virsh console sac
   ```

Replacing nodes

The following topics explain how to replace nodes:

- **Replacing failed nodes or system disks**
- **Scalable unit (SU) leader node operations**

For more information about other hardware operations and about replacing other types of cluster components, see the following:

HPE Performance Cluster Manager Administration Guide

Replacing failed nodes or system disks

This topic introduces the process for installing and configuring a spare node or for replacing failed system disks.

The failed node can be an admin, leader, or non-ICE 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. Verify that you have an appropriate spare node or spare system disks.

   A cold spare node is equivalent to 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.

   As part of maintaining the cluster, make sure that you always have the following two types of spare nodes:

   - One spare for the admin node.
   - One spare for a leader or non-ICE compute node.

   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 non-ICE compute nodes.

     For example, an admin node does not PXE boot by default. However, you can configure a leader node to PXE boot by default. 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.
   - Depending on your site policy, the management cards of an admin node might or might not be configured to use DHCP by default.

     The management cards of leader nodes and non-ICE compute nodes must be configured to use DHCP by default. Otherwise, attempts to use the `discover` command to configure the node into the cluster will fail.

2. Complete one of the following procedures:

- **Replacing an entire node**. Use this procedure if the entire node has failed.
- **Replacing failed system disks in a node**. Use this procedure if only the disks in the node have failed.
- **Replacing a node and reinstalling the original system disks**. Use this procedure if the node has failed, but the system disks in the node are functional.

NOTE: If you are using multiple root slots, the installation procedures affect only the current slot.

Replacing an entire node

Use the procedure in this topic to replace an entire node without preserving the original system disks.

Procedure

1. Connect a keyboard, video screen, and mouse to the node.
2. If possible, power down the failed node.
3. Examine and label the power cables.

   Before you disconnect any cables, make sure they are labeled. Make sure that you are familiar enough with the cabling to re-cable the new node at the end of this procedure.
4. Disconnect all power cables.
5. Unplug the Ethernet cables 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 cables must be connected to the same plugs on the cold spare unit.
6. Remove any peripheral components, such as a keyboard, video screen, or mouse, from the node.
7. Remove the failed node from the rack.
8. Install the shelf spare node into the rack.
9. Connect the Ethernet cables in the same way they were connected to the replaced node.
10. Connect AC power.
11. Connect to the node through the management card or attach a keyboard, video screen, and mouse to the node.
12. (Conditional) From the admin node, update the cluster manager database.

    Complete this step as follows:

    - If you replaced an ICE leader node, a scalable unit (SU) leader node, or a non-ICE compute node complete this step.
    - If you completed this procedure as part of the process to add an additional, new scalable unit (SU) leader node, do not complete this step. Instead, proceed to the following:

      Adding scalable unit (SU) leader nodes

    Update the following in the cluster database:

- The MAC address of the spare
- The MAC address of the management card 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. For more information about how to retrieve the MAC address of the spare, see the management card documentation for the spare. For example, see the iLO server guide for the spare.

From the admin node, query and set the MAC addresses in the database. The following table shows the command parameters that you can use:

Example 1. The following example displays the MAC address of non-ICE compute node `n0`:

```
# cm node show -M -n n0
NODE   NETWORK.NAME   IPADDRESS    SUBNETMASK    MACADDRESS
n0     None           172.23.0.3   255.255.0.0   00:25:90:fd:3c:28
```

NOTE: The preceding output has been truncated from the right for inclusion in this documentation.

Example 2. The following example sets the MAC address of `n0`:

```
# cm node set --mac-address 00:25:90:04:4e:01 -n n0
```

Example 3. The following example shows the MAC address of the management card on `n1`:

```
# cm node show -B -n n1
NODE                 CARDIPADDRESS        CARDMACADDRESS       CARDTYPE
PROTOCOL
n1                   172.24.0.11          00:25:90:cd:7d:83    IPMI
dcmi,ipmi
```

Example 4. The following example sets the MAC address of the management card on `n0`:

```
# cm node set --bmc-mac-address 00:25:90:03:51:1d -n n0
```

13. Power up the replaced node.

Replacing failed system disks in a node

The procedure in this topic explains how to replace system disks. The procedure assumes that the rest of the node is operating appropriately. You can reinstall the system disks into a replacement node.

Do not use this procedure to replace the Gluster disks in a scalable unit (SU) leader node. Use the following procedure for that task:

Replacing a Gluster disk

Procedure

1. Obtain new system disks from Hewlett Packard Enterprise.
2. Connect a keyboard, video screen, and mouse to the node.
3. Power down the node that contains the failed system disks.
4. Remove the failed system disks from the node.
5. Install the new system disks into the node.
6. Use the management card to connect to the failed node or attach a keyboard, video screen, and mouse to the failed node.

If your system disks were part of a RAID, use the RAID controller interface to configure the disks into a RAID. The RAID controller interface is often part of the BIOS. See the RAID documentation for the node.

7. Power up the node.
8. (Conditional) Configure the RAID controller to use the new system disks.
9. From the admin node, install the cluster manager software on the new system disks:

   ```
   cm node provision -n hostname -i image
   ```

   The variables are as follows:

   - For *hostname*, specify the hostname of the node with the new system disks.
   - For *image*, specify the name of the image that had been installed on the failed system disks.

Replacing a node and reinstalling the original system disks

Use the procedure in this topic if a node is no longer functioning, but the system disks within the node are still useful. This procedure explains the following:

- Removing good disks from a failed node
- Preserving the removed disks
- Installing the preserved disks from the failed node into a new node

Procedure

1. Connect a keyboard, video screen, and mouse to the node.
2. If possible, power down the failed node.
3. Examine and label the power cables.

   Before you disconnect any cables, make sure they are labeled. Make sure that you are familiar enough with the cabling to re-cable the new node at the end of this procedure.
4. Disconnect all power cables.
5. Unplug the Ethernet cables 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 cables must be connected to the same plugs on the cold spare unit.
6. Remove any peripheral components, such as a keyboard, video screen, or mouse, from the node.
7. Remove the failed node from the rack.
8. Remove the system disks from the failed node.

   That is, open the failed node and remove the system disks.
9. Remove the system disks from the new node.

   That is, pull the current system disks, using their carriers, and set the disks aside.
10. Insert the preserved disks from the failed node into the new node (the shelf spare).
11. Insert the new node, with the preserved disks, back into the rack.

12. Connect AC power to the new node.

13. Connect a keyboard, video screen, and mouse to the new node.

14. From the admin node, update the cluster manager database.

    Update the following in the cluster database:

    - The MAC address of the spare
    - The MAC address of the management card 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. For more information about how to retrieve the MAC address of the spare, see the management card documentation for the spare. For example, see the iLO server guide for the spare.

    From the admin node, query and set the MAC addresses in the database. The following table shows the command parameters that you can use:

    Example 1. The following example displays the MAC address of non-ICE compute node `n0`:

    ```
    # cm node show -M -n n0
    NODE  NETWORK.NAME  IPADDRESS   SUBNETMASK   MACADDRESS
    n0    None          172.23.0.3  255.255.0.0  00:25:90:fd:3c:28
    ```

    NOTE: The preceding output has been truncated from the right for inclusion in this documentation.

    Example 2. The following example sets the MAC address of `n0`:

    ```
    # cm node set --mac-address 00:25:90:04:4e:01 -n n0
    ```

    Example 3. The following example shows the MAC address of the BMC on `n1`:

    ```
    # cm node show -B -n n1
    NODE                 CARDIPADDRESS        CARDMACADDRESS       CARDTYPE
    PROTOCOL
    n1                   172.24.0.11          00:25:90:cd:7d:83    IPMI
    dcmi,ipmi
    ```

    Example 4. The following example sets the MAC address of the BMC on `n0`:

    ```
    # cm node set --bmc-mac-address 00:25:90:03:51:1d -n n0
    ```

15. Power up the replaced node.

16. (Conditional) Interrupt the boot-up sequence in BIOS and enter the RAID configuration tool.

    Complete this step if the disk or disks being replaced represent a RAID configuration.

    The RAID controller facilitates the importing of drives and volumes into the new node. After the RAID is configured, the node might reboot or you might have to reset the node. Typically, the node boots normally.

    For information, see the RAID documentation for the node.

Scalable unit (SU) leader node operations

Replacing a scalable unit (SU) leader node

Procedure

1. Take out the failing node.

Complete the following procedure:

Replacing an entire node

2. Log into the admin node as the root user.

3. Make sure that the replacement node is operational.

   For example, if the node is up, you can `ssh` to the node.

4. Open the `/opt/clmgr/etc/su-leader-nodes.lst` file and verify the Gluster disk LUN path.

   If necessary, correct the path.

   For information about how to edit this file, see the following:

   Obtaining or creating a scalable unit (SU) leader node list file

5. Enter the following command to configure the new SU leader node into the cluster:

   ```
   su-leader-setup --reintegrate-whole-leader [--destroy-gluster-disk] hostname
   ```

   The parameters are as follows:

   - The `--reintegrate-whole-leader` parameter is required.
   - The `--destroy-gluster-disk` parameter is optional. Use this parameter if there are partitions or information on the disk at this time. This parameter reformats the disk.
   - For *hostname*, specify the hostname of the new SU leader node.

6. Enter the following commands to monitor the Gluster rebalancing:

   - `ssh leaderX gluster volume heal cm_shared info summary`

     The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.

   - `ssh leaderX gluster volume heal cm_logs info summary`
   - `ssh leaderX gluster volume heal ctdb info summary`

   For *X*, specify the number of one of the leader nodes.

   For example, the biggest volume, `cm_shared`, consumes the most time. The other volumes heal more quickly. The following commands show how to monitor the `cm_shared` volume:

   ```
   leader1:~ # gluster volume heal cm_shared info summary
   Brick 172.23.0.3:/data/brick_cm_shared
   Status: Connected
   Total Number of entries: 6378
   Number of entries in heal pending: 6378
   Number of entries in split-brain: 0
   Number of entries possibly healing: 0

   Brick 172.23.0.4:/data/brick_cm_shared
   Status: Connected
   Total Number of entries: 10892
   Number of entries in heal pending: 10892
   Number of entries in split-brain: 0
   Number of entries possibly healing: 0
   ```

```
Brick 172.23.0.5:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.6:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.7:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.8:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.9:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.10:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.11:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0
```

When the healing is complete, the numbers for the entries are all 0. For example:

```
leader3:~ # gluster volume heal cm_shared info summary
Brick 172.23.0.3:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.4:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.5:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.6:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.7:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.8:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.9:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.10:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0

Brick 172.23.0.11:/data/brick_cm_shared
Status: Connected
Total Number of entries: 0
```

```
Number of entries in heal pending: 0
Number of entries in split-brain: 0
Number of entries possibly healing: 0
```

For information, see the Gluster documentation.

7. Back up the cluster configuration.

   At this time, continue to the following procedure to back up the cluster configuration files:

   Backing up the cluster

Adding scalable unit (SU) leader nodes

When you add SU leader nodes, always add a multiple of three nodes. For example, you can add three, six, or nine SU leader nodes at a time.

Prerequisites
The cluster is configured. You want to add SU leader nodes.

Procedure

1. For each new SU leader node, obtain the MAC address of each of the following:

   - The management card MAC address
   - The GbE/GigE MAC address

   For information about how to retrieve the MAC address of the spare, see the documentation for the management card for the spare.

2. Log into the admin node as the root user.

3. Update the cluster definition file that contains information about switches and SU leader nodes.

   Add the new nodes to the file. For each new node, include the MAC addresses for the nodes in the lines that define the nodes.

   The following example shows an updated file. The file includes information about the three new SU leader nodes at the end:

```
# File mgmtsw_suleader.config
# Cluster definition file for management switches and SU leader nodes on an HPE apollo cluster
[templates]
name=su-leader, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, mgmt_net_interfaces="eno1,eno2",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=802.3ad, redundant_mgmt_network=yes,
switch_mgmt_network=yes, transport=udpcast, tpm_boot=no, dhcp_bootfile=grub2, disk_bootloader=no,
predictable_net_names=yes, console_device=ttyS0, conserver_ondemand=no, conserver_logging=yes,
rootfs=disk, card_type=iLO, baud_rate=115200,
force_disk="/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:0", su_leader_role=yes
[nic_templates]
template=su-leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="eno1,eno2"
template=su-leader, network=head-bmc, net_ifs="bmc0"
template=su-leader, network=ib-0, net_ifs="ib0"
template=su-leader, network=ib-1, net_ifs="ib1"
[discover]
internal_name=mgmtsw0, mgmt_net_macs="40:b9:3c:a2:54:50", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine, mgmt_net_ip=172.23.255.254
internal_name=mgmtsw1, mgmt_net_macs="40:b9:3c:a4:6c:a7", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.1
internal_name=mgmtsw2, mgmt_net_macs="40:b9:3c:a6:6a:a2", mgmt_net_name=head,
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.100.2
internal_name=service1, hostname1=leader1, mgmt_bmc_net_macs="20:67:7c:e4:8a:8a",
mgmt_net_macs="00:0f:53:21:98:30,00:0f:53:21:98:31", mgmt_net_ip=172.23.10.1,
mgmt_bmc_net_ip=172.24.10.1, template_name=su-leader
internal_name=service2, hostname1=leader2, mgmt_bmc_net_macs="20:67:7c:e4:9a:ba",
mgmt_net_macs="00:0f:53:21:98:90,00:0f:53:21:98:91", mgmt_net_ip=172.23.10.2,
mgmt_bmc_net_ip=172.24.10.2, template_name=su-leader
internal_name=service3, hostname1=leader3, mgmt_bmc_net_macs="20:67:7c:e4:8a:7a",
mgmt_net_macs="00:0f:53:3c:e0:a0,00:0f:53:3c:e0:a1", mgmt_net_ip=172.23.10.3,
mgmt_bmc_net_ip=172.24.10.3, template_name=su-leader
# New SU leader nodes:
# The MAC address for the second management MAC can appear in the mgmt_net_macs= field.
# It is not required.
# The cluster manager detects the second MAC address when the node boots.
internal_name=service251, hostname1=leader4, mgmt_bmc_net_macs="20:67:6c:e4:8a:2a",
mgmt_net_macs="00:0f:53:20:98:30", template_name=su-leader
internal_name=service252, hostname1=leader5, mgmt_bmc_net_macs="20:69:7c:e5:9a:ba",
mgmt_net_macs="00:0f:53:21:98:90", template_name=su-leader
internal_name=service253, hostname1=leader6, mgmt_bmc_net_macs="20:67:7c:e8:8a:4c",
mgmt_net_macs="00:0f:53:3c:e0:a0", template_name=su-leader
```

Notice the specifications for the new nodes in the preceding file.

4. Run the `discover` command to configure the new SU leader nodes into the cluster.

   For example, to configure the nodes described in `mgmtsw_suleader.config`, enter the following command:

   ```
   # discover --configfile mgmtsw_suleader.config --node 251 --node 252 \
   --node 253
   ```

5. Wait for the new nodes to come up.

   For example, if the node is up, you can `ssh` to the node.

6. Open the `/opt/clmgr/etc/su-leader-nodes.lst` file and add entries for the new SU leader nodes.

   For information about how to edit this file, see the following:

   Obtaining or creating a scalable unit (SU) leader node list file

7. Use the `su-leader-setup` command to configure the new leader nodes.

   The format for the command is as follows:

   ```
   su-leader-setup --add-leaders new_hostname1,new_hostname2,new_hostname3
   ```

   For *new_hostname1,new_hostname2,new_hostnamethree*, specify the hostnames of the three new SU leader nodes.

   For example:

   ```
   # su-leader-setup --add-leaders leader4,leader5,leader6
   ```

8. Enter the following commands to verify that there are Gluster volumes for all the SU leader nodes:

   - `ssh leaderX gluster volume heal cm_shared info summary`

     The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.

   - `ssh leaderX gluster volume heal cm_logs info summary`
   - `ssh leaderX gluster volume heal ctdb info summary`

   For *X*, specify the number of one of the leader nodes.

   For command examples, see the following:

   Replacing a scalable unit (SU) leader node

9. Back up the cluster configuration.

   At this time, continue to the following procedure to back up the cluster configuration files:

Backing up the cluster

Replacing a Gluster disk

The following procedure explains how to replace a failed Gluster disk on a scalable unit (SU) leader node. Typically, the Gluster disk is a set of disks in a RAID. Complete this procedure when the following conditions occur:

- The file system becomes corrupted.
- The disk fails.
- The RAID fails.

Procedure

1. Remove the failed Gluster disk and install a new Gluster disk into the SU leader node.
2. Open the `/opt/clmgr/etc/su-leader-nodes.lst` file and verify the Gluster disk LUN path.

   If necessary, correct the path.

   For information about how to edit this file, see the following:

   Obtaining or creating a scalable unit (SU) leader node list file
3. Use the `su-leader-setup` command to enable the cluster manager to recognize the new disk.

   The format for this command is as follows:

   ```
   su-leader-setup --replace-failed-brick hostname
   ```

   For *hostname*, specify the SU leader hostname.
4. Enter the following commands to verify the healing status:

   - `ssh leaderX gluster volume heal cm_shared info summary`

     The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.
   - `ssh leaderX gluster volume heal cm_logs info summary`
   - `ssh leaderX gluster volume heal ctdb info summary`

   For *X*, specify the number of one of the leader nodes.

   For command examples, see the following:

   Replacing a scalable unit (SU) leader node

Migrating from the HPE Cluster Management Utility (CMU) or the SGI Management Suite (SMC)

The following topics explain how to migrate to the HPE Performance Cluster Manager:

- Use the following procedure to migrate from HPE Cluster Management Utility:

  Migrating from the HPE Cluster Management Utility (CMU)

- Use the following procedure to migrate from SGI Management Suite (also known as SGI Management Center):

  Migrating from SGI Management Suite

NOTE: The migration procedures do not describe an **upgrade**. These procedures describe a **migration**.

For information about how to upgrade to a newer version of the HPE Performance Cluster Manager, see the following:

Upgrading from an HPE Performance Cluster Manager 1.x release

Migrating from the HPE Cluster Management Utility (CMU)

Before you migrate from CMU to the new HPE Performance Cluster Manager, be aware of the following new features and requirements:

- HPE Performance Cluster Manager supports migration from compute node images that run one of the following operating system levels:

  - CentOS 7.4 or later
  - RHEL 7.4 or later
  - SLES 12 SP3 or later

  The HPE Performance Cluster Manager can import, manage, and deploy compute node images that run these operating system distributions. In addition, the HPE Performance Cluster Manager can create new compute node images based on these operating system distributions.

- Relative to CMU, the HPE Performance Cluster Manager stores more cluster information in its internal database. The additional information enables more cluster configuration automation.

  For example, to configure TCP-over-InfiniBand, CMU requires scripting the creation of the node-specific `ifcfg-ibX` files in a post-cloning script. With the HPE Performance Cluster Manager, you only configure the following in the cluster definition file:

  - One or two InfiniBand networks.
  - (Optionally) One or two specific IP addresses for each node.

  The cluster manager configures the `ifcfg-ibX` files for you during deployment.

- The migration process requires you to install an RPM on your CMU admin node that contains a script. When the script runs, it creates an HPE Performance Cluster Manager configuration file called `config.txt`. This configuration file is based on your current cluster configuration. Be sure to review `config.txt` before you install the HPE Performance Cluster Manager because you might need to update the `config.txt` file. The `config.txt` file resides in the directory from which you started the migration script.

- Image management is different between CMU and the HPE Performance Cluster Manager.

In CMU, compute node images are built on one compute node. When the image is ready for distribution to each node, CMU performs the following actions:

    - It archives and compresses the image into a large file.
    - It stores the new file on the admin node.
    - It clones out the file to other compute nodes.

  When using the HPE Performance Cluster Manager, compute node images are created locally on the admin node, in a directory, and then deployed to the compute nodes. When migrating from CMU to the HPE Performance Cluster Manager, you can import your CMU images into the HPE Performance Cluster Manager. After the import, deploy them using the HPE Performance Cluster Manager. The images are required to contain an operating system that the HPE Performance Cluster Manager supports.

- The HPE Performance Cluster Manager supports **slots** for all nodes in the cluster. This feature divides up the local disks in each node into slots and installs a complete operating system into each slot. By default, the cluster manager creates two slots: one slot for the current operating system and another slot for a new (or different) operating system. The slot management feature automatically defines and manages partitions for each operating system installation. If you require a specific partition layout for your compute nodes, you can configure custom partitions.

  For more information about slots, see the following:

  About slots

  For more information about custom partitions, see the following:

  (Conditional) Configuring custom partitions on the admin node

- The migration process includes reinstalling the admin node. CMU calls this component the **head node**. If you have other software on that node, either migrate that software to another server or archive and restore that software back onto the admin node.

The migration consists of the following procedures:

- **Verifying and preparing the environment**
- **Installing the new software**
- **Completing the migration**

Verifying and preparing the environment

Procedure

1. Verify the CMU level that is running on the cluster.

   The migration requires CMU 8.2.4, which is the latest CMU release available.

2. Review the information in the following manual to familiarize yourself with HPCM, the hardware environment, and networking information:
   HPE Performance Cluster Manager Getting Started Guide

3. Download the migration software RPM from the HPE support center, and write the RPM to a directory on the admin node.

   The support center is at the following link:

   https://support.hpe.com/hpesc/public/home

   If the cluster has a high availability (HA) admin node, write the RPM to the primary admin node.

4. From the admin node, enter the following command to install the RPM:

   ```
   # rpm -ivh migrate-cm*.rpm
   ```

5. Log into the head node as the root user and ensure the following:

   - All the compute nodes are up, running, and accessible by the root user through passwordless `ssh`.
   - The iLO credentials are cached correctly in CMU.

6. Enter the following command to start the migration process:

   ```
   # hpcm_migrate_cmu
   ```

   This command gathers details about the existing CMU cluster. In addition, this command creates a `config.txt` file that can be used to build out the cluster when managed by the HPE Performance Cluster Manager. This command does not change the existing CMU cluster. The `config.txt` file resides in the directory from which you entered the command in this step.

7. Review the content of the following, newly generated file and complete any tasks prescribed in the file:

   `config.txt`

   The following are some notes on how to review and update the `config.txt` file:

   - The `public` network definition must match the external connection to the admin node.
   - The `head` network definition must match the expected management network.
   - The `head-bmc` network is the interface that the admin node uses to access the BMCs of the compute nodes. CMU does not store the BMC netmask, so the `hpcm_migrate_cmu` command tries to calculate the BMC netmask from the known BMC IP addresses.
   - The `admin house` interface must be the network interface card (NIC) that is connected to the external site network.
   - The `mgmt_net_interfaces` must specify the NIC (or NICs for bonding) that is connected to the cluster management network.
   - The `predictable_net_names` attribute controls how the HPE Performance Cluster Manager configures the network interface names on the cluster, including the admin node, as follows:

     - When set to `yes` (default), the cluster manager uses the configured interface names for each node.
     - When set to `no`, the cluster manager uses `eth`*`X`* for all network interfaces.

     Be sure that this setting and the interface names match the expected behavior. For example, if the admin server was installed and/or configured with `biosdevname=1` as a kernel parameter, then set `predictable_net_names=no`. The `biosdevname=1` kernel parameter declares that all network interfaces are named `eth`*`X`*.

     Most default installations have enabled predictable network interface names, so the default setting of `predictable_net_names=yes` is correct. Notice that, if necessary, the `predictable_net_names` parameter can be set on a per-node basis. For an example, see the following:

     Retrieving predictable NIC names by examining the cluster definition file

   If any of these settings are not accurate, correct them before proceeding. Inaccurate settings cause the cluster configuration step to fail.

8. (Conditional) Prepare the HA head node for removal of the CMU RPM.

Complete this step if the cluster is configured for high availability.

On the primary node, enter the following commands:

```
# service cmu unset_audit
# /opt/cmu/tools/saveConfig.tcl -p root -c /opt/cmu-store
```

9. Enter the following commands to remove the CMU RPM:

```
# service cmu stop
# systemctl disable cmu
# rpm -e cmu
```

Use the `rpm -e` command to remove any other add-ons.

For example, if the cluster has the `cmu-arm64-moonshot-addon` for Arm (AArch64) support, remove that. The new cluster manager does not support the Windows operating system. If the CMU cluster included the Windows Moonshot add-on package, remove that, too. To remove both of these add-on packages, enter the following command:

```
# rpm -e cmu cmu-arm64-moonshot-addon cmu-windows-moonshot-addon
```

10. Archive and save the `/opt/cmu` directory to another server or storage device outside of the cluster.

On an HA cluster, this step does the following:

- It saves the shared `/opt/cmu-store` directory to another server or storage device.
- It assumes that you want to use the device that hosts `/opt/cmu-store` as the storage device for the HPE Performance Cluster Manager HA cluster.

For example, the following commands create a `tar` file of the `/opt/cmu` directory and copy the `tar` file to another server, as follows:

```
# cd /opt
# tar zcf cmu.tgz cmu
# scp cmu.tgz other_storage_server
```

Depending on the number of images, the `tar` command in the preceding example can take a while to complete.

11. (Conditional) Archive and save any other services you want to preserve to another server or storage device outside of the cluster.

The HPE Performance Cluster Manager provides a central way to manage repositories. To preserve your autoinstall repositories, make sure to save the elements required to recreate those directories on the new admin server. For example, save the operating system distribution ISO file because the HPE Performance Cluster Manager can manage the ISO file after the migration. Also, if the admin node hosts Altair PBS Works, Slurm, or other services, archive (or save) them to an off-cluster storage location.

12. Proceed to the following:

Installing the new software

Installing the new software

Procedure

1. Complete the following procedure:

Preparing to install the operating system and the cluster manager jointly

Notes:

- Step **1** has already been done. You can skip this step.
- Step **8** refers to the config.txt file. You already created this file, so you can skip this step.

2. (Optional) Complete the following procedure:

   (Conditional) Configuring custom partitions on the admin node

3. Complete the following procedure:

   Inserting the installation DVD and booting the system

4. Complete one of the following procedures:

   - **Configuring RHEL 8.X or RHEL 7.X on the admin node**
   - **Configuring SLES 15 SPX and SLES 12 SPX on the admin node**

5. (Optional) Complete the procedures in the following chapter:

   Configuring a high availability (HA) admin node

6. Locate the cluster manager software distribution DVDs, or verify the path to the online software repository at your site.

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

7. From the VGA screen, or through an `ssh` connection, log into the admin node as the root user, as follows:

   - For a single-node admin node, Hewlett Packard Enterprise recommends that you run the cluster configuration tool as follows:

     - From the VGA screen

       or

     - From an `ssh` session to the admin node

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

   - For an HA admin node, create an `ssh` connection to the host that is running the `virt` resource, and enter the `virt-viewer` command. For example:

     ```
     # ssh -C -XY root@phys_admin1
     # virt-viewer
     ```

     If the Virtual Machine Manager interface does not appear, log into the physical node, and enter `virt-viewer sac` on that host.

8. (Conditional) Open the ports that the cluster manager requires.

   Complete this step if you configured a firewall on the admin node or anywhere else in the cluster.

   The cluster manager requires the following ports:

   - External ports required for SSH: TCP 22
   - (Conditional) External port required for Kibana: TCP 5601

     TCP port 5601 is required if you want to use Kibana to access the centralized log files.

- External ports required for webpage and GUI: TCP 80, 443, 1099, and 49150.

  It is possible to start the cluster manager web server on a different port. For more information, see the following:

  Starting the cluster manager web server on a non-default port

- Internal ports required for monitoring: UDP 48555 - 49587

To avoid monitoring, UDPcast, and other failures, do not permit other software to use the cluster manager ports.

For more information about port requirements, see the following:

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

9. Restore `config.txt` and `/opt/cmu` to the admin node.

   For a CMU HA cluster, rename the `/opt/cmu-store` directory to `/opt/cmu`.

10. Compare the admin interfaces and other configuration settings configured in the `config.txt` file on the admin node.

    The admin interfaces and other configuration settings in `config.txt` must match the interfaces and settings after the fresh installation.

    Correct settings as needed.

11. Proceed to the following:

    Completing the migration

    NOTE: In **Completing the migration**, the first three steps might already have been completed. Verify these steps.

Completing the migration

Procedure

1. Verify the domain and reset if needed.

   If an alternative domain is not found, the cluster manager sets a default domain of `pcm-cluster.net` in `config.txt`.

   If necessary, open the `/etc/hosts` file in a text editor and add a line in the following format to set the domain:

   `ip_addr fqdn hostname`

2. Run the following command to create the operating system distribution image repository from the ISO that is included as part of the larger ISO:

   ```
   cm repo add distro_iso
   ```

   For *distro_iso*, specify the path to the operating system ISO.

3. Run the following command to create the cluster manager image repository:

   ```
   # cm repo add /mnt/cm-*
   ```

4. Enter the following command to configure the cluster:

   ```
   # configure-cluster --configfile=config.txt
   ```

5. (Conditional) Import the node images.

   Complete this step if the node images include the operating system levels that the HPE Performance Cluster Manager supports.

For more information about image management, see the following:

HPE Performance Cluster Manager Administration Guide

For each node image that you want to include in the new cluster manager, complete the following steps:

a. Use the following command to create a soft link:

```
ln -s /opt/cmu/image/image_name /opt/clmgr/image/image_name
```

For *image_name*, specify the name of one software image.

b. Run the following command to see if the matching cluster manager repository exists and is selected:

```
# cm repo show
```

If the output does not show the repository present and selected, use one or both of the following:

- Use `cm repo add` to add the cluster manager repository.
- Use `cm repo select` to select the cluster manager repository.

c. Run the following command to see if the matching operating system repository exists and is selected:

```
# cm repo show
```

If the output does not show the repository present and selected, use one or both of the following:

- Use `cm repo add` to add the operating system repository.
- Use `cm repo select` to select the operating system repository.

d. Verify that the correct cluster manager and operating system repositories are present and are selected.

Run the `cm repo show` command once more. When correct, only those images that match the image that you want to import appear in the output as selected.

If nonmatching repositories for the cluster manager or for the operating system appear as selected in the output, use the `cm repo unselect` command to clear them.

e. Use the following command to import the image:

```
cm image create -i image_name --use-existing
```

f. After the import is complete, use the following command to remove the soft link:

```
rm -f /opt/clmgr/image/image_name
```

g. If you have more images, repeat the command sequence in this step for another image.

6. Enter the following command to complete the cluster configuration:

```
# discover --configfile config.txt --all --skip-provision
```

7. (Optional) Assign images to nodes.

By default, the cluster manager creates and assigns a stock image to all nodes. This image matches the operating system distribution of the admin node.

To change this default behavior so that images that were successfully imported are assigned to nodes, run the following commands:

a. Run the following command, and note the kernel version for the image that you want to assign to a node:

```
# cm image show -d
```

b. Run the following command to assign the given image and kernel to the given node:

```
cinstallman --assign-image --image name --node node --kernel version
```

c. Run the `cinstallman` command in the following format:

```
cinstallman –next-boot image –node name
```

This `cinstallman` command specifies to install the assigned image on the node the next time the node boots.

For information about the `cinstallman` command, see the following:

- The `cinstallman` manpage or help output
- **HPE Performance Cluster Manager Administration Guide**

8. Restart monitoring.

   HPCM monitoring is similar to legacy CMU monitoring. To restart monitoring, complete the following steps:

   - On the top menu, click **Options** > **Enter Admin mode**
   - Use one of the following methods to recreate the network groups:
     - Within the GUI, click **Cluster Administration** > **Network Group Management**
     - From the command line, use the `cmu_add_network_group` and `cmu_change_network_group` commands.
   - On the top menu, click **Monitoring** > **Restart Monitoring Engine**.

9. Reboot the cluster, and verify that all nodes booted.

10. Rebuild the cluster compute node images, and install the new images on the compute nodes.

    Use the image management commands in the following:

    HPE Performance Cluster Manager Administration Guide

Migrating from SGI Management Suite

Prerequisites

The SGI Management Suite was also known as SGI Management Center, or SMC.

The procedure in this topic applies to clusters with and without a highly available admin node. For systems with a highly available admin node, note the following modifications:

- When you migrate, you migrate the virtual machine admin node that runs on the physical admin nodes. Complete the following steps on the virtual machine admin node to save information from the virtual machine:

Step **1** through Step **10**.

- When completing Step **11**, install the HPE Performance Cluster Manager on the two physical admin nodes and on the virtual machine. On the two physical admin nodes, use the same IP address and configuration that you used when they ran SMC. On the virtual machine admin node, use the information that was saved.

The procedure in this topic describes how to migrate an SMC cluster to the new HPE Performance Cluster Manager.

Procedure

1. Log into the admin node as the root user.
2. Verify the SGI Management Suite release level that is running on the cluster.

   The migration requires the cluster to run the latest SGI Management Suite available, which is release 3.5.
3. Verify the admin node operating system level.

   The migration software requires one of the following operating systems on the admin node:

   - RHEL 7.4 or later
   - CentOS 7.4 or later
   - SLES 12 SP3 or later
4. Verify the operating system level for each node image and update the images as needed.

   The migration software requires the node images to be at one of the following operating system levels:

   - RHEL 7.4 or later
   - CentOS 7.4 or later
   - SLES 12 SP3 or later

   Proceed as follows:

   - If the node images are at the levels in the preceding list, proceed to the following:

     Step **8**
   - If the node images are older than the node images in the preceding list, proceed to the following:

     Step **5**
5. (Conditional) Retrieve the operating system software for the image, and use the `crepo` command to add the software repositories.

   Complete this step only if you have to upgrade the operating system software.

   Example 1. If your images are for RHEL, add the RHEL 7.7 media:

   ```
   # crepo --add RHEL-7.7-20190723.1-Server-x86_64-dvd1.iso
   # crepo --unselect old_distro_repo
   # crepo --select Red-Hat-Enterprise-Linux-7.7-x86_64
   ```

   Example 2. If your images are for SLES, add the SLES 12 SP4 media:

   ```
   # crepo --add SLE-12-SP4-Server-DVD-x86_64-GM-DVD1.iso
   # crepo --unselect old_distro_repo
   # crepo --select SUSE-Linux-Enterprise-Server-12-SP4
   ```

6. (Conditional) Use the `--update-image` parameter and `--image` parameter on the `cinstallman` command to install updated packages.

   Complete this step only if you have to upgrade the operating system software.

   For example, to install RHEL 7.7 packages, enter the following:

   ```
   # cinstallman --update-image --image ice-compute-rhel7.7
   # cinstallman --update-image --image rhel7.7
   # cinstallman --update-image --image lead-rhel7.7
   ```

   Proceed as follows:

   - If the `cinstallman` commands were successful, proceed to the following:

     Step **8**

   - If the `cinstallman` commands were not successful, proceed to the following:

     Step **7**

7. (Conditional) Use operating-system-specific commands to install updated packages.

   Complete this step only if you have to upgrade the operating system software.

   For example:

   - For RHEL:

     Use the `--yum-image` parameter and the `--image` parameter to update the image.

     The format is as follows:

     ```
     cinstallman --yum-image --image image_name yum_params
     cinstallman --yum-image --image image_name yum_params
     cinstallman --yum-image --image image_name yum_params
     ```

     For *image_name*, specify one of the cluster image names. For example:

     ```
     ice-compute-rhel7.7
     rhel7.7
     lead-rhel7.7
     ```

     For *yum_params*, specify one or more `yum` command parameters. Enclose the parameters in quotation marks (`" "`) if you specify any flags.

     For more information, see the RHEL documentation.

   - For SLES:

     Use `--zypper-image` parameter and the `--image` parameter to update the image.

     The format is as follows:

     ```
     cinstallman --zypper-image --image image_name zypper_params
     cinstallman --zypper-image --image image_name zypper_params
     cinstallman --zypper-image --image image_name zypper_params
     ```

     For *image_name*, specify one of the cluster image names. For example:

```
ice-compute-sles12sp4
sles12sp4
lead-sles12sp4
```

For *zypper_params*, specify one or more `zypper` command parameters. Enclose the parameters in quotation marks (`"  "`) if you specify any flags. For more information, see the SLES documentation.

8. Download the migration software RPM from the HPE support center, and write the RPM to a directory on the admin node:

The support center is at the following link:

https://support.hpe.com/hpesc/public/home

9. From the admin node, enter the following command to install the RPM:

```
# rpm -ivh migrate-cm*.rpm
```

10. Enter the following command to run the migration software and to create a set of cluster definition files:

```
# /opt/clmgr/bin/hpcm_migrate_smc --backup [--mount-path path]
```

The `--mount_path path` parameter writes a tar file that includes the images, the cluster definition file, and the slot map files to the *path* directory. Use this parameter as follows:

- For clusters with two or more slots, the `--mount_path` parameter is optional.
- For clusters with only one slot, the `--mount_path` parameter is required. This parameter specifies the location to which the migration software writes all your cluster information.

  For *path*, specify the full path to a directory outside the cluster.

By default, this command writes the following:

- A set of cluster definition files to the following directory:

  `/var/backups/migration_backup/configfiles`

- A set of slot map files to the following directory:

  `/var/backups/migration_backup/slotmaps`

- A dump of the cluster database in the following directory:

  `/var/backups/`*`node_hostname`*`_db_backup.sq/`

- If you specify a `--mount-path` parameter, the command writes information to a tar file in the following directory on the remote system:

  `/var/backups/`*`admin_hostname`*`_backup.tar.gz`

  The tar file contains the following:

  - The current cluster definition file
  - The current slot map files
  - The current images

11. Reinstall the cluster manager in a new slot.

Complete the following procedures:

- **Installing the operating system and the cluster manager jointly**
- (Conditional) **Configuring a high availability (HA) admin node**

12. (Conditional) Create custom repositories.

    Complete this step if you have repositories for software that was not developed by HPE or SGI.

    For example, if you have a repository for Slurm, create a repository for Slurm. Use the instructions in the following:

    HPE Performance Cluster Manager Administration Guide

13. Use one of the following methods to extract the tar file:

    - Method 1

      Use the following command if you specified a mount path on another system when you entered the `hpcm_migrate_smc --backup` command:

      ```
      # hpcm_migrate_smc --import --mount-path path
      ```

      For *path*, specify the full path to the mount point.

    - Method 2

      Use the following command if you did not write the backup files to another system:

      ```
      # hpcm_migrate_smc --import --slot num
      ```

      For *num*, specify the slot that hosts the installation that you backed up.

      The commands in this step extract the tar file to the following directory, either on another system or on the current slot:

      ```
      /var/backups/migration_backup
      ```

14. Use the `cd` command to change to the following directory:

    ```
    /var/backups/migration_backup/configfiles
    ```

15. Choose one of the cluster definition files, and update that file as necessary.

    The migration script created a set of configuration files. Hewlett Packard Enterprise recommends that you use the default file for the migration, but you can use any of them. The following table shows the configuration files:

Type	Information	Use if ...
admin_hostname`_backup_default.config`	BMC and IP (default)	You want to use new node images. This file does not include image information for any node. Hewlett Packard Enterprise recommends that you use this file.
admin_hostname`_backup_image_info.config`	BMC and image	You want to use imported images and new IP addresses.
admin_hostname`_backup_all_info.config`	BMC, IP, and image	You want to use imported images and old IP addresses.
admin_hostname`_backup_basic.config`	BMC	You want to use new images and new IP addresses.

For the preceding configuration files, the information included is as follows:

- The BMC information includes username and password information.
- The image information includes the images assigned to each node, including the kernel version and all kernel parameters.

For example, on a cluster that has the hostname of `mycluster`, the following cluster definition files exist:

```
mycluster_backup_all_info.config
mycluster_backup_basic_info.config
mycluster_backup_default.config
mycluster_backup_image_info.config
```

For example, to specify new InfiniBand networks or to specify new admin node attributes, add them at this time. To add them, edit the configuration file you want to use during the migration.

16. Enter the following command to start the cluster configuration tool:

```
configure-cluster --configfile /var/backups/migration_backup/configfiles/
configfile
```

For *configfile*, specify the name of the cluster definition file you want to use. Select one file from the table in the preceding step.

17. Enter the following command to import the node images:

```
hpcm_migrate_smc --import-images [--keep-image-names]
```

This step writes the node images to the cluster database and to the following location:

```
/opt/clmgr/image/images
```

By default, this command renames the migrated images and adds the following prefix:

`smc-`

The prefix indicates that the image was migrated from SMC.

For example, the default version of this command changes a default flat compute node image name of `sles12sp4` to `smc-sles12sp4`.

When you specify `--keep-image-names`, the image names are not renamed. Use this parameter if you want to retain the SMC image names in your HPE Performance Cluster Manager configuration. If you use this parameter, the result is that the migration script replaces images with other images that have the same name.

18. Update the migrated node images.

    HPCM supports specific operating system levels in node images. Complete the following steps:

    - Run the following command, verify the output, and ensure that the correct repositories are selected:

      ```
      # cm repo show
      ```

    - Run the following command, verify the output, and ensure that the correct repository groups are selected:

      ```
      # cm repo group show
      ```

      The preceding command checks for the presence of repository groups. It is possible that the cluster does not have any repository groups.

    - Run the following command:

      ```
      hpcm_migrate_smc \
      --update-image "image_name [image_name] ... [image_name]" \
      [--repo-group group_name]
      ```

      The variables are as follows:

      - For *image_name*, specify one or more image names. Use a space to separate each image name, and enclose the list of image names in quotation marks (`"  "`).
      - For *group_name*, specify a repository group.

19. Run the `discover` command.

    Use one of the following methods:

    - Method 1 - Use new images (recommended)

      Use the `discover` command in the following format:

      ```
      discover --configfile /path_to_new_configfile --all
      ```

      For *path_to_new_configfile*, specify the location of the cluster configuration file. This file is named with the hostname of the admin node appended with `_backup_default.config`. For example, `mycluster_backup_default.config`.

      This method uses the default configuration file without the image information or kernel information. The `discover` command uses the images you created recently with the cluster configuration tool.

For example:

```
# discover --configfile \
/var/backups/migration_backup/configfiles/ \
mycluster_backup_default.config --all
```

- Method 2 - Use imported SMC images

  For this method, the HPE Performance Cluster Manager adds discover command uses the imported images when it runs. That is, the migration process uses the old, imported images and associates the nodes with those images. Use the discover command in the following format:

  ```
  discover --configfile /path_to_new_configfile --all
  ```

  For *path_to_new_configfile*, specify the location of the cluster configuration file. This file is named with the hostname of the admin node appended with -all_info.config.

  For example:

  ```
  # discover --configfile \
  /var/backups/migration_backup/configfiles/mycluster-all_info.config \
  --all
  ```

20. Reboot the cluster, and verify that all nodes booted.

Upgrading from an HPE Performance Cluster Manager 1.x release

Starting the upgrade

Procedure

1. Log into the admin node as the root user.
2. Back up the current cluster installation.

   Clone the current slot to a new, target slot. For example, assume that you are on slot 1, and slot 2 is open:

   ```
   # clone-slot --source 1 --dest 2
   ```

   Alternatively, complete the following procedure:

   Backing up the cluster

3. List the services that are running:

   ```
   # systemctl list-units --type=service --state=running > services.file
   ```

   Save `services.file` to another system at your site. After the upgrade, you can compare the content of this file to the list of services running after the upgrade.

4. (Conditional) Remove the Elasticsearch, Kibana, and Logstash RPMs.

   Complete this step to upgrade from HPE Performance Cluster Manager 1.1 or 1.2 to HPE Performance Cluster Manager 1.3.1. You do not need to complete this step to upgrade from HPE Performance Cluster Manager 1.3 to HPE Performance Cluster Manager 1.3.1.

   Enter the following command:

   ```
   # rpm -e kibana elasticsearch logstash --noscripts
   ```

5. (Conditional) Remove the Logstash directory.

   Complete this step to upgrade from HPE Performance Cluster Manager 1.1 or 1.2 to HPE Performance Cluster Manager 1.3.1. You do not need to complete this step to upgrade from HPE Performance Cluster Manager 1.3 to HPE Performance Cluster Manager 1.3.1.

   Enter the following command:

   ```
   # rm -rf /usr/share/logstash
   ```

6. Complete the following steps to add new repositories:

   a. Add the cluster manager software:

   ```
   # crepo --add cm-1.3.1-*.iso
   ```

   b. Use the `crepo` command in the following format to add the operating system software:

   ```
   crepo --add new_distro_media.iso
   ```

For *new_distro_media*, specify the `.iso` file for the new operating system software. For example:

```
# crepo --add RHEL-7.7-20190723.1-Server-x86_64-dvd1.iso
```

c. (Conditional) Add MPI software:

```
# crepo --add hpe-mpi-1.6*.iso
```

7. Create repository groups for any repositories you created.

For example:

```
# crepo --add-group hpcm-1.3.1 Red-Hat-Enterprise-Linux-7.7-x86_64 \
Cluster-Manager-1.3.1-rhel77-x86_64 \
HPE-MPI-1.6-rhel77-x86_64
```

8. Upgrade the admin node.

This step differs depending on the admin node operating system.

- For an admin node that runs SLES, enter the following:

```
# cinstallman --zypper-node --node admin --repo-group hpcm-1.3.1 \
"dup --allow-vendor-change"
```

- For an admin node that runs RHEL 8.X or CentOS 8.X, enter the following commands:

```
# cinstallman --dnf-node --node admin \
--repo-group hpcm-1.3.1 update yume
# cinstallman --update-node --node admin --repo-group hpcm-1.3.1
```

- For an admin node that runs RHEL 7.X or CentOS 7.X, enter the following commands:

```
# cinstallman --yum-node --node admin \
--repo-group hpcm-1.3.1 update yume
# cinstallman --update-node --node admin --repo-group hpcm-1.3.1
```

NOTE: At this point in the upgrade process, the HPE Performance Cluster Manager 1.3.1 command set is available. For continuity purposes, however, the upgrade instructions show the command set that was available in previous releases.

9. Refresh the admin node RPM list.

For example, the following command removes an RPM list in a repository group:

```
# crepo --recreate-rpmlists
```

10. Refresh the admin node.

For example:

```
# cinstallman --refresh-node --node admin --repo-group hpcm-1.3.1 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.3.1-admin.rpmlist
```

11. Run the following script:

```
# /opt/sgi/lib/cluster-configuration
```

12. Run the following command:

```
# update-configs
```

13. (Conditional) Ensure that the CMU service can start.

    Complete this step for upgrades from HPE Performance Cluster Manager 1.3 to HPE Performance Cluster Manager 1.3.1. You do not need to complete this step to upgrade from HPE Performance Cluster Manager 1.0, 1.1, or 1.2 to HPE Performance Cluster Manager 1.3.1.

    Complete the following steps:

    a. Retrieve the IP address of the head interface:

       ```
       # cadmin --show-ips --node admin | grep "head"
       admin          172.23.0.1       head
       admin-bmc      172.24.0.1       head-bmc
       ```

       The preceding output shows the IP address of the head interface to be `172.23.0.1`.

    b. Open the following file in a text editor:

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

    c. Search for `CMU_CLUSTER_IP`, and ensure that `CMU_CLUSTER_IP` is set to the IP address of the head interface.

       For example, edit the line to appear as follows:

       ```
       CMU_CLUSTER_IP=172.23.0.1
       ```

    d. Restart the CMU service:

       ```
       # systemctl restart cmu.service
       ```

14. Reboot the admin node.

15. Wait for the admin node to boot the operating system.

16. Back up the images to VCS.

    For example, enter a `cinstallman` command in the following format for each node image:

    ```
    cinstallman --commit --image image \
    --msg "Image backup for hpcm 1.3.1 upgrade"
    ```

    For *image*, specify the name of one of the node images.

17. Proceed to the following:

    Upgrading the non-ICE compute nodes

Upgrading the non-ICE compute nodes

Procedure

1. Upgrade the non-ICE compute node image.

   This step differs depending on the operating system on the non-ICE compute nodes.

- For a SLES non-ICE compute node image, enter a `cinstallman` command in the following format:

  ```
  cinstallman --zypper-image --image image --repo-group hpcm-1.3.1 \
  --duk "dup --allow-vendor-change"
  ```

- For a RHEL 8.X or CentOS 8.X non-ICE compute node image, enter commands in the following format:

  ```
  cinstallman --dnf-image --image image \
  --repo-group hpcm-1.3.1 update yume --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

- For a RHEL 7.X or CentOS 7.X non-ICE compute node image, enter commands in the following format:

  ```
  cinstallman --yum-image --image image \
  --repo-group hpcm-1.3.1 update yume --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

For *image*, specify the name of one of the non-ICE compute node images.

2. Complete the following steps to refresh the non-ICE compute node image:

a. Enter two `cinstallman` commands in the following formats:

```
cinstallman --refresh-image --image image \
--repo-group hpcm-1.3.1 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.3.1.rpmlist \
--duk

cinstallman --update-kernels --image image
```

b. Enter the following `cinstallman` command to update the miniroot:

```
cinstallman --update-miniroot --image image --recreate \
--repo-group hpcm-1.3.1
```

3. Upgrade the non-ICE compute nodes.

This step differs depending on the operating system of the non-ICE compute nodes.

- For SLES non-ICE compute nodes, enter the following command:

  ```
  # cinstallman --zypper-node --node "n*" --repo-group hpcm-1.3.1 "dup --allow-vendor-change"
  ```

- For RHEL 8.X or CentOS 8.X non-ICE compute nodes, enter the following commands:

  ```
  # cinstallman --dnf-node --node "n*" --repo-group hpcm-1.3.1 update yume
  ```

and

```
# cinstallman --update-node --node "n*" --repo-group hpcm-1.3.1
```

- For RHEL 7.X or CentOS 7.X non-ICE compute nodes, enter the following commands:

  ```
  # cinstallman --yum-node --node "n*" --repo-group hpcm-1.3.1 update yume
  ```

  ```
  # cinstallman --update-node --node "n*" --repo-group hpcm-1.3.1
  ```

4. Complete the following steps to refresh the non-ICE compute nodes and reboot them:

   a. Refresh the non-ICE compute node images.

      For example:

      ```
      # cinstallman --refresh-node --node "n*" --repo-group hpcm-1.3.1 \
      --rpmlist /opt/clmgr/image/rpmlists/generated/\
      generated-group-hpcm-1.3.1.rpmlist
      ```

   b. (Optional) Update the kernels.

      Use the `cinstallman` command in the following format:

      ```
      cinstallman --assign-image --image image --kernel new_kernel \
      --node node ... node
      ```

   c. Use the following `cpower` command to reboot the non-ICE compute nodes:

      ```
      cpower node reboot node
      ```

5. Proceed to one of the following:

   - If the cluster has ICE leader nodes, proceed to the following:

     (Conditional) Upgrading the ICE leader nodes and the ICE compute nodes

   - If the cluster does not have ICE leader nodes, proceed to the following:

     Completing the upgrade

(Conditional) Upgrading the ICE leader nodes and the ICE compute nodes

Complete this procedure if the cluster contains ICE leader nodes.

Procedure

1. Upgrade the ICE leader node images.

   This step differs depending on the ICE leader node operating system.

- For ICE leader node images that runs SLES, enter a `cinstallman` command in the following format:

  ```
  cinstallman --zypper-image --image image --repo-group hpcm-1.3.1 \
  --duk "dup --allow-vendor-change"
  ```

- For a RHEL 8.X or CentOS 8.X ICE leader node image, enter commands in the following format:

  ```
  cinstallman --dnf-image --image image \
  --repo-group hpcm-1.3.1 update yume --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

- For a RHEL 7.X or CentOS 7.X ICE leader node image, enter commands in the following format:

  ```
  cinstallman --yum-image --image image \
  --repo-group hpcm-1.3.1 update yume --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

For *image*, specify the name of one of the ICE leader node images.

2. Complete the following steps to refresh the ICE leader node image:

a. Enter two `cinstallman` commands in the following formats:

```
cinstallman --refresh-image --image image \
--repo-group hpcm-1.3.1 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.3.1-lead.rpmlist \
--duk

cinstallman --update-kernels --image image
```

b. Enter the following `cinstallman` command to update the miniroot:

```
cinstallman --update-miniroot --image image --recreate \
--repo-group hpcm-1.3.1
```

3. Upgrade the ICE leader nodes.

This step differs depending on the ICE leader node operating system.

- If the ICE leader nodes run SLES, enter the following:

  ```
  # cinstallman --zypper-node --node "r*lead" --repo-group hpcm-1.3.1
  ```

- If the ICE leader nodes run RHEL 8.X or CentOS 8.X, enter the following:

  ```
  # cinstallman --dnf-node --node "r*lead" \
  --repo-group hpcm-1.3.1 update yume
  ```

and

```
# cinstallman --update-node --node "r*lead" --repo-group hpcm-1.3.1
```

- If the ICE leader nodes run RHEL 7.X or CentOS 7.X, enter the following:

  ```
  # cinstallman --yum-node --node "r*lead" \
  --repo-group hpcm-1.3.1 update yume
  ```

  and

  ```
  # cinstallman --update-node --node "r*lead" --repo-group hpcm-1.3.1
  ```

4. Complete the following steps to refresh the ICE leader nodes and reboot them.

a. Refresh the ICE leader node images.

For example:

```
# cinstallman --refresh-node --node "r*lead" --repo-group hpcm-1.3.1 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.3.1-lead.rpmlist
```

b. (Optional) Update the kernels.

Use the `cinstallman` command in the following format:

```
cinstallman --assign-image --image image --kernel new_kernel --node node ... node
```

c. Use the following `cpower` command to reboot the ICE leader nodes:

```
# cpower leader reboot "r*lead*
```

5. Upgrade the ICE compute node image.

This step differs depending on the operating system:

- For a SLES ICE compute node image, enter a `cinstallman` command in the following format:

  ```
  cinstallman --zypper-image --image image --repo-group hpcm-1.3.1 \
  --duk "dup --allow-vendor-change"
  ```

- For a RHEL 8.X or CentOS 8.X ICE compute node image, enter `cinstallman` commands in the following format:

  ```
  cinstallman --dnf-image --image image --repo-group hpcm-1.3.1 update yume \
  --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

- For a RHEL 7.X or CentOS 7.X ICE compute node image, enter `cinstallman` commands in the following format:

  ```
  cinstallman --yum-image --image image --repo-group hpcm-1.3.1 update yume \
  --duk
  ```

  and

  ```
  cinstallman --update-image --image image --repo-group hpcm-1.3.1 --duk
  ```

6. Complete the following steps to refresh the ICE compute node image:

a. Enter two `cinstallman` commands in the following formats:

```
cinstallman --refresh-image --image image \
--repo-group hpcm-1.3.1 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.3.1-ice.rpmlist \
--duk

cinstallman --update-kernels --image image
```

b. Enter the following `cinstallman` command to update the miniroot:

```
cinstallman --update-miniroot --image image --recreate \
--repo-group hpcm-1.3.1
```

7. Complete the following steps to push upgraded images to the ICE leader nodes:

a. Shut down the ICE compute nodes:

```
# cpower node shutdown "r*i*n*"
```

b. Use the `cimage` command in the following format to push the images to the ICE compute nodes:

```
cimage --push-rack ice-image "r*"
```

c. (Optional) Assign new kernels to the ICE compute node images. Enter the following command:

```
cimage --set ice_image new_kernel ice_node
```

For *ice_node*, specify the image name for the ICE compute node image.

For *new_kernel*, specify the kernel associated with the ICE compute node image.

For *node_name*, specify the ICE compute node to receive the new kernel and image. The *node_name* accepts globbing.

d. Power up the ICE compute nodes:

```
# cpower node on "r*i*n*"
```

8. Proceed to the following:

Completing the upgrade

Completing the upgrade

Procedure

1. (Optional) Enable Monitoring.

The upgrade process disables monitoring. The commands for re-enabling and restarting monitoring are as follows:

```
cm monitoring monitor enable
cm monitoring monitor start
```

For *monitor*, enter one of the following:

- alerta
- elk
- ganglia
- kafka
- nagios
- native

For example, enter the following commands to re-enable and restart Ganglia:

```
# cm monitoring ganglia enable
# cm monitoring ganglia start
```

2. Enter the following command to list the services that are running:

```
# systemctl list-units --type=service --state=running
```

Compare the output from this command to the services.file content that you saved at the beginning of this procedure.

Websites

General websites

Hewlett Packard Enterprise Information Library

https://www.hpe.com/info/EIL

Single Point of Connectivity Knowledge (SPOCK) Storage compatibility matrix

https://www.hpe.com/storage/spock

Storage white papers and analyst reports

https://www.hpe.com/storage/whitepapers

For additional websites, see **Support and other resources**.

Support and other resources

Accessing Hewlett Packard Enterprise Support

- For live assistance, go to the Contact Hewlett Packard Enterprise Worldwide website:

  https://www.hpe.com/info/assistance

- To access documentation and support services, go to the Hewlett Packard Enterprise Support Center website:

  https://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

  https://www.hpe.com/support/hpesc

  Hewlett Packard Enterprise Support Center: Software downloads

  https://www.hpe.com/support/downloads

  My HPE Software Center

  https://www.hpe.com/software/hpesoftwarecenter

- To subscribe to eNewsletters and alerts:

  https://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:

  https://www.hpe.com/support/AccessToSupportMaterials

ⓘ **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.

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

https://www.hpe.com/services/getconnected

HPE Proactive Care services

https://www.hpe.com/services/proactivecare

HPE Datacenter Care services

https://www.hpe.com/services/datacentercare

HPE Proactive Care service: Supported products list

https://www.hpe.com/services/proactivecaresupportedproducts

HPE Proactive Care advanced service: Supported products list

https://www.hpe.com/services/proactivecareadvancedsupportedproducts

Proactive Care customer information

Proactive Care central

https://www.hpe.com/services/proactivecarecentral

Proactive Care service activation

https://www.hpe.com/services/proactivecarecentralgetstarted

Warranty information

To view the warranty information for your product, see the links provided below:

HPE ProLiant and IA-32 Servers and Options

https://www.hpe.com/support/ProLiantServers-Warranties

HPE Enterprise and Cloudline Servers

https://www.hpe.com/support/EnterpriseServers-Warranties

HPE Storage Products

https://www.hpe.com/support/Storage-Warranties

HPE Networking Products

https://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:

https://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:

https://www.hpe.com/info/reach

For Hewlett Packard Enterprise product environmental and safety information and compliance data, including RoHS and REACH, see:

https://www.hpe.com/info/ecodata

For Hewlett Packard Enterprise environmental information, including company programs, product recycling, and energy efficiency, see:

https://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**). 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.

YaST navigation

The following table shows SLES YaST navigation key sequences.

Key	Action
Tab **Alt** + **Tab** **Esc** + **Tab** **Shift** + **Tab**	Moves you from label to label or from list to list.
Ctrl + **L**	Refreshes the screen.
Enter	Starts a module from a selected category, runs an action, or activates a menu item.
Up arrow	Changes the category. Selects the next category up.
Down arrow	Changes the category. Selects the next category down.
Right arrow	Starts a module from the selected category.
Shift + **right arrow** **Ctrl** + **A**	Scrolls horizontally to the right. Useful in screens if use of the **left arrow** key would otherwise change the active pane or current selection list.
Alt + *letter* **Esc** + *letter*	Selects the label or action that begins with the *letter* you select. Labels and selected fields in the display contain a highlighted *letter*.
Exit	Quits the YaST interface.

Subnetwork information

Cluster hardware components can be connected to multiple networks. Generally, a network is assigned to a single subnet.

A **subnet** is a logical subdivision of an IP network. A subnet keeps broadcast traffic from the various hosts within the subnet contained in its own subnet. This action helps clusters to scale properly. Additionally, if layer-3 IP routing is not configured, the components that reside in a subnet can communicate only with other components within the subnet.

The cluster management software uses a variety of networking concepts to accomplish the architecture design goals for various cluster types. These concepts include the following:

- Virtual Local Area Network (VLAN / 802.1Q) tagging
- Supernetting
- Layer 3 IP routing
- Subinterfaces

Network and subnet information within a cluster

Table 4: Network and subnet information shows the following for the networks that the cluster management software uses:

- The names of the components on the networks
- The default allocation of the systemwide IP address ranges on the networks

The following notes pertain to these networks:

- Generally, a node and its management card (iLO or BMC) reside in the same VLAN. However, these components do not reside in the same IP subnet range. This separation prevents cross-communication between the host node and its management interface.
- Systems with HPE SGI 8600 ICE liquid-cooled hardware contain cooling distribution unit (CDU) and cooling rack controller (CRC) components. These components have static IP addresses assigned to them. In addition, they have to be put into the Mcell VLAN manually as an untagged port. The Mcell VLAN has a default 802.1Q tag of `3`.

  The admin node uses a Linux 802.1Q-tagged interface under the `bond0` interface to communicate to these CDU/CRC components. Therefore, the cluster requires that the Mcell 802.1Q tag be allowed on the ports connected to the admin node.

  Several components on HPE SGI 8600 clusters exist in their own, respective rack 802.1Q-tagged VLAN. This practice allows their respective leader nodes to communicate only to the chassis management controllers (CMCs) and compute nodes in its own rack. These components are as follows:

  - ICE CMCs
  - ICE InfiniBand switches
  - ICE compute nodes

  The following are additional notes regarding these components:

- If layer-3 routing is configured and enabled, the admin node can communicate to these ICE components directly through routing.
- ICE chassis management controllers (CMCs) communicate with CDU/CRC components (when present). For that reason, the ports connected to ICE CMCs pass the Mcell 802.1Q tag as a tagged port.
- By default, the rack 802.1Q-tagged VLAN uses the following equation:

  rack# + 100

  For example:

  Rack 1 = (1 + 100) = VLAN 101

  Rack 2 = (2 + 100) = VLAN 102

  Rack 150 = (150 + 100) = VLAN 250

  .

  .

  .

- HPE Apollo 9000 clusters contain CMCs, HPE Adaptive Rack Cooling Systems (ARCS) components, or both.

  All components behind a CMC take on the network/VLAN settings of the CMC. An HPE Apollo 9000 CMC can have InfiniBand switches, non-ICE compute nodes, CDUs, ARCS components, and PDUs behind it.

  By default, two networks are automatically generated for every eight CMCs that are attached to a cluster. These two networks exist under a single 802.1Q VLAN beginning at number 2001 and ending at 2999. The naming scheme for these two networks is as follows:

  - `hostmgmt`*`XXXX`* - for host management traffic such as PXE, SSH, TFTP, and ICMP
  - `hostctrl`*`XXXX`* - for control traffic such as the management card, Redfish, SNMP, IPMI, and power

  If the automatic generation of networks is disabled, CMCs and their respective components end up in the `head` and `head-bmc` networks. To disable the automatic generation of networks, before you run the `discover` command to configure any component, run the following command:

  ```
  # cadmin --set-cmcs-per-mgmt-vlan 0
  ```

  To verify the number of CMCs per management VLAN that is allowed, enter the following command:

  ```
  # cadmin --show-cmcs-per-mgmt-vlan
  ```

  HPE Apollo clusters do not use the MCell network.

- Generally, a cluster component is connected to a management switch that is contained within the management network. Each switchport to which a component is connected has at least one VLAN assigned to it. However, each port could be assigned more than one VLAN. This technology is known as **VLAN tagging**.

  Table 4: Network and subnet information shows the components in a given VLAN as either an **untagged port** or a **tagged port**.

  The following are additional notes:

  - At a minimum, all switchports are put into a VLAN as an untagged port. This is also known as a **native VLAN** or a **default VLAN** in some networking nomenclature.
  - All traffic coming from a component that is otherwise untagged is put into an untagged VLAN.
  - A switchport is not required to allow tagged VLANs.

Some switchports allow a tagged VLAN. These switchports forward the traffic coming out of the VLAN when the traffic coming out of a component with a VLAN tag matches the switchport configuration.

- A switchport can allow zero, one, two, or many tagged VLANs at the same time.

Table 4: Network and subnet information

VLAN #	Subnet Name	IP Range / Subnet Mask	Nodes in Subnet
1	head	172.23.0.0/16	Admin bond0 (untagged) ICE leader (untagged) Generic compute (untagged) Scalable unit (SU) leader (untagged) Management switch VLAN 1 (untagged)
1	head-bmc	172.24.0.0/16	Admin bond0:bmc (untagged) ICE leader management card (untagged) SU leader management card (untagged) Generic compute management card (untagged) Cooling devices such as ARCS and CDUs (untagged)
3	mcell-net	172.26.0.0/16	Admin (tagged) ICE CMCs (tagged) CDUs (untagged) CRCs (untagged) Management switch interswitch Links (tagged)
101~1100 Example 1 Example 2	vlan*XXX*:gbe vlan101:gbe vlan102:gbe	10.159.X.X/22 10.159.0.0/22 10.159.4.0/22	ICE compute (untagged) ICE leader (tagged)
101~1100 Example 1 Example 2	vlan*XXX*:bmc vlan101:bmc vlan102:bmc	10.160.X.X/22 10.160.0.0/22 10.160.4.0/22	ICE compute (untagged) ICE CMC (untagged) ICE InfiniBand switch (untagged) ICE leader (tagged)

Table Continued

VLAN #	Subnet Name	IP Range / Subnet Mask	Nodes in Subnet
2001~2999	`hostmgmt`*`XXXX`*	`10.170.X.X/22`	HPE Apollo 9000 compute (untagged)
Example 1	`hostmgmt2001`	`10.170.0.0/22`	
Example 2	`hostmgmt2002`	`10.170.4.0/22`	
2001~2999	`hostctrl`*`XXXX`*	`10.171.X.X/22`	HPE Apollo 9000 compute management card (untagged) HPE Apollo 9000 CMC (untagged) HPE Apollo 9000 InfiniBand switch (untagged) HPE Apollo 9000 CDU (untagged)
Example 1	`hostctrl2001`	`10.171.0.0/22`	
Example 2	`hostctrl2002`	`10.171.4.0/22`	
N/A	`ib-0`	`10.148.0.0/16`	Any component with InfiniBand interfaces
N/A	`ib-1`	`10.149.0.0/16`	Any component with InfiniBand interfaces

Naming conventions

The cluster management software has the following default naming formats for various components found within a cluster:

- If a component is configured into the cluster by using the `discover` command, you can specify a custom hostname of your choice.

  If you do not specify a hostname, the cluster management software uses the default naming convention.

- If a component is automatically added to the database, the naming convention is predetermined and cannot be specified at this time.

In a cluster definition file, for any given component, you can append the following parameter to assign a custom hostname to any component:

```
hostname1=hostname
```

For example, an entry for a non-ICE compute node might look like this:

```
internal_name=service1,mgmt_net_name=head,hostname1=r01n01,...
```

In the preceding example, the ellipsis (`...`) at the end represents the fact that you could specify many other configuration attributes on this line.

Table 5: Naming conventions includes information about various components, default naming conventions for each component, and the type of cluster in which these components can be found. The variables *T*, *X*, *Y*, and *Z* are always positive integer numbers. The examples represent the hostnames that can be seen once a component is added to the cluster management software.

Table 5: Naming conventions

Component	Internal name format	Examples	Found In

Admin node	*admin*	`myadmin`
		`sleet`
		`snow`
			All clusters

Ethernet management switch	`mgmtsw`*X*	`mgmtsw0` (spine) `mgmtsw1` (leaf)	All clusters
Ethernet data switch	`datasw`*X*	`datasw0` (spine) `datasw1` (leaf)	Clusters with an Ethernet high-speed fabric
InfiniBand data switch	`ibsw`*X*	`ibsw0` `ibsw1`	Clusters with an InfiniBand high-speed fabric
Non-ICE compute node	`service`*X*	`service1` `service100`	Clusters with generic compute resources
Scalable unit (SU) leader node	`leader`*X*	`leader1` `leader9`	Clusters with SU leader nodes
ICE leader (sometimes called a rack leader controller (RLC))	`r`*X*`lead`	`r1lead` `r10lead`	Clusters with HPE SGI 8600 ICE hardware
ICE compute node (nonadjustable)	`r`*X*`i`*Y*`n`*Z*	`r1i1n1` `r9i5n10`	Clusters with HPE SGI 8600 ICE hardware
ICE InfiniBand switch (nonadjustable)	`r`*X*`i`*Y*`s`*Z*`-bmc`	`r1i1s1-bmc`	Clusters with HPE SGI 8600 ICE hardware
ICE chassis management controller (CMC) (nonadjustable)	`r`*X*`i`*Y*`c`	`r1i0c` `r15i5c`	Clusters with HPE SGI 8600 ICE hardware

Management card interfaces	`service`*X*`-bmc`
	`leader`*X*`-bmc`
	`r`*X*`lead-bmc`
	`r`*X*`i`*Y*`n`*Z*`-bmc`
		`service1-bmc`
		`leader1-bmc`
		`r1lead-bmc`
		`r1i1n1-bmc`
			Clusters that contain components that have a management card. Management cards can be of an iLO or BMC interface type.

Table Continued

Component	Internal name format	Examples	Found In
HPE Apollo 9000 CMC (nonadjustable)	r*X*c*Y*	r1c1 r4c4	Clusters with HPE Apollo 9000 hardware
HPE Apollo 9000 non-ICE compute node	r*X*c*Y*t*T*n*Z*	r1c1t2n1 r2c1t1n0	Clusters with HPE Apollo 9000 hardware
Power distribution unit (PDU)	pdu*X*	pdu0 pdu1	Clusters with PDU hardware
HPE Adaptive Rack Cooling System (ARCS) device	cooldev*X*	cooldev0, cooldev1	HPE Apollo clusters
Cooling distribution units (CDUs)	cooldev*X*	cooldev0, cooldev1	HPE Apollo 9000 clusters

Liquid cooling cell network IP addresses (HPE SGI 8600 clusters)

To troubleshoot the liquid cooling cell cooling equipment on a hierarchical cluster, you must know the IP addresses of the following:

- The cooling rack controllers (CRCs)
- The cooling distribution units (CDUs)

NOTE: The information in this appendix section does not apply to HPE Apollo platforms. The HPE Apollo 9000 cluster does not restrict the cooling components to any particular IP address range.

To `ping` the component, the IP address is required. Each piece of equipment bears a label with its equipment number, as follows:

- For CRCs, the IP address is 172.26.128.*number*.
- For CDUs, the IP address is 172.26.144.*number*.

NOTE: To change the IP addresses for the CRCs and CDUs, contact your technical support representative.

The following table shows the IP addresses for CRCs and CDUs.

Table 6: Liquid cooling cell network associations

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
1	1	172.26.128.1	172.26.144.1
2	1	172.26.128.1	172.26.144.1
3	2	172.26.128.2	172.26.144.1
4	2	172.26.128.2	172.26.144.1
5	3	172.26.128.3	172.26.144.2
6	3	172.26.128.3	172.26.144.2
7	4	172.26.128.4	172.26.144.2
8	4	172.26.128.4	172.26.144.2
9	5	172.26.128.5	172.26.144.3

Table Continued

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
10	5	172.26.128.5	172.26.144.3
11	6	172.26.128.6	172.26.144.3
12	6	172.26.128.6	172.26.144.3
13	7	172.26.128.7	172.26.144.4
14	7	172.26.128.7	172.26.144.4
15	8	172.26.128.8	172.26.144.4
16	8	172.26.128.8	172.26.144.4
17	9	172.26.128.9	172.26.144.5
18	9	172.26.128.9	172.26.144.5
19	10	172.26.128.10	172.26.144.5
20	10	172.26.128.10	172.26.144.5
21	11	172.26.128.11	172.26.144.6
22	11	172.26.128.11	172.26.144.6
23	12	172.26.128.12	172.26.144.6
24	12	172.26.128.12	172.26.144.6
25	13	172.26.128.13	172.26.144.7
26	13	172.26.128.13	172.26.144.7
27	14	172.26.128.14	172.26.144.7
28	14	172.26.128.14	172.26.144.7
29	15	172.26.128.15	172.26.144.8
30	15	172.26.128.15	172.26.144.8
31	16	172.26.128.16	172.26.144.8

Table Continued

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
32	16	172.26.128.16	172.26.144.8
33	17	172.26.128.17	172.26.144.9
34	17	172.26.128.17	172.26.144.9
35	18	172.26.128.18	172.26.144.9
36	18	172.26.128.18	172.26.144.9
37	19	172.26.128.19	172.26.144.10
38	19	172.26.128.19	172.26.144.10
39	20	172.26.128.20	172.26.144.10
40	20	172.26.128.20	172.26.144.10
41	21	172.26.128.21	172.26.144.11
42	21	172.26.128.21	172.26.144.11
43	22	172.26.128.22	172.26.144.11
44	22	172.26.128.22	172.26.144.11
45	23	172.26.128.23	172.26.144.12
46	23	172.26.128.23	172.26.144.12
47	24	172.26.128.24	172.26.144.12
48	24	172.26.128.24	172.26.144.12
49	25	172.26.128.25	172.26.144.13
50	25	172.26.128.25	172.26.144.13
51	26	172.26.128.26	172.26.144.13
52	26	172.26.128.26	172.26.144.13
53	27	172.26.128.27	172.26.144.14

Table Continued

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
54	27	172.26.128.27	172.26.144.14
55	28	172.26.128.28	172.26.144.14
56	28	172.26.128.28	172.26.144.14
57	29	172.26.128.29	172.26.144.15
58	29	172.26.128.29	172.26.144.15
59	30	172.26.128.30	172.26.144.15
60	30	172.26.128.30	172.26.144.15
61	31	172.26.128.31	172.26.144.16
62	31	172.26.128.31	172.26.144.16
63	32	172.26.128.32	172.26.144.16
64	32	172.26.128.32	172.26.144.16
65	33	172.26.128.33	172.26.144.17
66	33	172.26.128.33	172.26.144.17
67	34	172.26.128.34	172.26.144.17
68	34	172.26.128.34	172.26.144.17
69	35	172.26.128.35	172.26.144.18
70	35	172.26.128.35	172.26.144.18
71	36	172.26.128.36	172.26.144.18
72	36	172.26.128.36	172.26.144.18
73	37	172.26.128.37	172.26.144.19
74	37	172.26.128.37	172.26.144.19
75	38	172.26.128.38	172.26.144.19

Table Continued

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
76	38	172.26.128.38	172.26.144.19
77	39	172.26.128.39	172.26.144.20
78	39	172.26.128.39	172.26.144.20
79	40	172.26.128.40	172.26.144.20
80	40	172.26.128.40	172.26.144.20
81	41	172.26.128.41	172.26.144.21
82	41	172.26.128.41	172.26.144.21
83	42	172.26.128.42	172.26.144.21
84	42	172.26.128.42	172.26.144.21
85	43	172.26.128.43	172.26.144.22
86	43	172.26.128.43	172.26.144.22
87	44	172.26.128.44	172.26.144.22
88	44	172.26.128.44	172.26.144.22
89	45	172.26.128.45	172.26.144.23
90	45	172.26.128.45	172.26.144.23
91	46	172.26.128.46	172.26.144.23
92	46	172.26.128.46	172.26.144.23
93	47	172.26.128.47	172.26.144.24
94	47	172.26.128.47	172.26.144.24
95	48	172.26.128.48	172.26.144.24
96	48	172.26.128.48	172.26.144.24
97	49	172.26.128.49	172.26.144.25

Table Continued

Physical rack number	Logical rack number	Cooling rack controllers (CRCs)	Cooling distribution unit (CDUs)
98	49	172.26.128.49	172.26.144.25
99	50	172.26.128.50	172.26.144.25
100	50	172.26.128.50	172.26.144.25

Default partition layout information

The default partition layout uses the GUID partition table (GPT) and the GRUB version 2 boot system. Alternatively, to create a custom partitioning scheme for the cluster, see the following:

(Conditional) Configuring custom partitions on the admin node

Partition layout for a one-slot cluster

Table 7: Partition layout for a single-boot cluster shows the partition layout for a one-slot cluster. This layout yields one boot partition. If you configure a single-slot system and later decide to add another partition, the addition process destroys all the data on your system.

Table 7: Partition layout for a single-boot cluster

Partition	File system type	File system label	Notes
1	Ext4	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 data for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext4	`sgiboot`	Slot 1 `/boot` partition.
12-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	Notice that the `/boot/efi` partition is used only on systems with UEFI BIOS.
22-30	N/A	N/A	N/A
31	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot`	Slot 1 `/` partition.

Partition layout for a two-slot cluster

Table 8: Partition layout for a dual-boot cluster shows the partition layout for a two-slot cluster. This layout yields two boot partitions.

Table 8: Partition layout for a dual-boot cluster

Partition	File system type	File system label	Notes
1	Ext4	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 data for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext4	`sgiboot`	Slot 1 `/boot` partition.
12	Ext4	`sgiboot2`	Slot 2 `/boot` partition.
13-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	Slot 1 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
22	VFAT	`sgiefi2`	Slot 2 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
23-30	N/A	N/A	N/A
31	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot`	Slot 1 / partition.
32	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot2`	Slot 2 / partition.

Partition layout for a five-slot cluster

Table 9: Partition layout for a quintuple-boot cluster shows the partition layout for a five-slot cluster. This layout yields five boot partitions.

Table 9: Partition layout for a quintuple-boot cluster

Partition	File system type	File system label	Notes
1	Ext4	`sgidata`	Contains slot information. On the admin node, contains GRUB version 2 for choosing root slots at boot time.
2	swap	`sgiswap`	Swap partition.
3-10	N/A	N/A	N/A
11	Ext4	`sgiboot`	Slot 1 `/boot` partition.
12	Ext4	`sgiboot2`	Slot 2 `/boot` partition.
13	Ext4	`sgiboot3`	Slot 3 `/boot` partition.
14	Ext4	`sgiboot4`	Slot 4 `/boot` partition.
15	Ext4	`sgiboot5`	Slot 5 `/boot` partition.
16-20	N/A	N/A	N/A
21	VFAT	`sgiefi`	Slot 1 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
22	VFAT	`sgiefi2`	Slot 2 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
23	VFAT	`sgiefi3`	Slot 3 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.

Table Continued

Partition	File system type	File system label	Notes
24	VFAT	`sgiefi4`	Slot 4 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
25	VFAT	`sgiefi5`	Slot 5 `/boot/efi` partition. EFI BIOS clusters only. On x86_64 BIOS clusters, this partition is unused.
26-30	N/A	N/A	N/A
31	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot`	Slot 1 / partition.
32	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot2`	Slot 2 / partition.
33	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot3`	Slot 3 / partition.
34	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot4`	Slot 4 / partition.
35	Ext4 on the admin node and on non-ICE compute nodes. XFS on leader nodes.	`sgiroot5`	Slot 5 / partition.

Specifying configuration attributes

You can specify cluster configuration information several ways. For example:

- When you configure the cluster for the first time, you can provide configuration information as follows:
  - In the cluster definition file
  - As responses to prompts from the online cluster configuration tool
- When you add nodes to a cluster, you can specify node attributes as parameters to the `discover` command that you use to configure the nodes.
- When you use the `cadmin` command or the `cm node set` command, you set and apply an attribute.
- When you use the `cattr` command, you set an attribute.

The cluster manager supports several global cluster attributes. Some attributes can be specified in more than one way. The documentation for each attribute includes the following information:

- A description of the attribute
- The attribute default value
- Other valid values or ranges of values
- The commands or files that you can use to specify the value

NOTE: The configuration attributes in the topics that follow appear as specified in one of the following:

- The cluster definition file
- The `sac-ha-initial-setup.conf` file

In many cases, you can set or clear these attributes by using commands such as `cadmin` or `cm node set`. On a command line, the attribute specification often replaces underscore characters (_) with hyphens (-). For example, you can set the UDPcast attribute `udpcast_max_bitrate` in the cluster definition file. However, on the `cm node set` command, the format is `udpcast-max-bitrate`. For more information, see the manpages for the individual commands.

Provisioning options

`image`

Specifies the image for a node.

Values = The name of the image.

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

kernel

Specifies the kernel for a node.

Values = The version of the kernel.

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

nfs_writable_type

Specifies the type of writable area for NFS root file systems. Only valid when `rootfs=nfs` is in effect. For more information, see the `cinstallman`(1) manpage.

Values = `nfs-overmount`, `nfs-overlay`, `tmpfs-overmount`, or `tmpfs-overlay`.

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

rootfs

Sets the root file system type for a node. For more information, see the `cinstallman`(1) manpage.

Values = `disk`, `tmpfs`, `nfs`, `custom`

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

tpm_boot

Enables the node to boot, or not, as a trusted platform module (TPM).

Values = yes or no

Default = no

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

transport

Sets the image transport method.

Values = rsync, bt, udpcast

Default = rsync

Range = N/A

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

UDPcast options

edns_udp_size

Specifies the edns-udp-size option in /etc/named.conf. This value is the default packet size, in bytes, that remote servers can receive.

Default = 512.

Values = any positive integer number.

Accepted by:

cattr command

udpcast_max_bitrate

Specifies the maximum numbers of bits that are conveyed or processed per second. This attribute is expressed as a number followed by a unit of measure, such as m.

Default = 900m.

Values = any positive integer number followed by a unit of measure. The default unit of measure is m (megabytes). For the list of units of measure, see the udp-sender(1) manpage.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- cm node set command
- discover command

udpcast_max_wait

Specifies the greatest amount of time that can elapse between when the first client node connects and any other client nodes connect. Clients that connect after this time has elapsed receive their software in a subsequent broadcast.

Default = 10.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- cm node set command
- discover command

udpcast_mcast_rdv_addr

Specifies the UDPcast rendezvous (RDV) address. Used for senders and receivers to find each other.

The admin node default address and the global (leader node) default address are different. If you change the global setting, which is used by leaders, also make the following changes:

- Adjust `--set-udpcast-mcast-rdv-addr`.
- Use the `cimage` command to push an image and initiate changes on the leader nodes.

Default for the admin node = `239.0.0.1`.

Default for the global (leaders) = `224.0.0.1`.

Values = any valid IP address.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

`udpcast_min_receivers`

Specifies the minimum number of receiver nodes for UDPcast.

Default = `1`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

`udpcast_min_wait`

Specifies the minimum amount of time that the system waits, while allowing clients to connect, before the software broadcast begins. This specification is the time between when the first client node connects and any other client nodes connect. The UDPcast distributes the software to all clients that connect during this interval.

Default = `10`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command

- `cm node set` command
- `discover` command

udpcast_rexmit_hello_interval

Specifies the frequency with which the UDP sender transmits `hello` packets.

NOTE: The admin node has a different default than the leader nodes.

The defaults are as follows:

- For the admin node, the default is `5000` (5 seconds).
- For the leader nodes, the default is `0`.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

The `--rexmit-hello-interval` setting is especially important when the rendezvous (RDV) address is not 224.0.0.1. 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 that the Ethernet switch detects. The Ethernet switch then updates its tables with this information, thus allowing the multicast packets to properly route through the switch. The problem is that sometimes the UDP receiver sends its connection packet before the switch has had a chance to update the switch routing. If the request packet is not detected by the UDP sender on the admin node, the UDP receiver could wait forever for a UDPcast stream. For example, the sender might not detect the packet because the packet was sent before the switch was set up to pass the packet.

The `udpcast_rexmit_hello_interval` value configures the UDP sender to send a `HELLO` packet at regular intervals and configures UDP receivers to respond to the packet. This way, even if the UDP receiver request is missed, the UDP receiver sends a fresh request after seeing a `HELLO` packet from the UDP sender.

By default, HPE sets the `udpcast_rexmit_hello_interval` value to 5000 (5 seconds) for UDP senders running on the admin node

By default, on leader nodes, the UDP senders are set to 0 (disabled). Typically, an interval is not needed when the following conditions both exist:

- When 224.0.0.1 is the RDV address
- When there are no VLANs being crossed

To change the RDV address used by leader nodes to serve `tmpfs` ICE compute nodes, set the `udpcast_rexmit_hello_interval` value to a site-specific value. This action lets you avoid the situation described in this topic. To change the value, use the following command:

```
cm node set -g --udpcast-rexmit-hello-interval value
```

To display this value, use the following command:

```
# cadmin --show-udpcast-rexmit-hello-interval
```

The leader nodes use the global value when serving `tmpfs` ICE compute nodes. The admin node uses its value when it serves the following node types when using the UDPcast transport mechanism:

- Leader nodes
- Non-ICE compute nodes

For more information, see the information about the `--rexmit-hello-interval` on the `udp-sender` manpage.

`udpcast_ttl`

Sets the UDPcast time to live (TTL), which specifies the number of VLAN boundaries a request can cross.

NOTE: The admin node has a different default than the leader nodes.

The defaults are as follows:

- For the admin node, the default is `2`. The admin nodes serve the leader nodes and the non-ICE compute nodes.
- For leader nodes, the default is `1`. Leader nodes serve only the nodes under their control.

When `udpcast_ttl=1`, the request cannot cross a VLAN boundary. When `udpcast_ttl=2`, the request can cross one VLAN boundary. If your site has routed management networks, a data transmission might have to cross from one VLAN to another. If your site has no routed management networks, or if your site policy requires, you can set `udpcast_ttl=1` for both the leader nodes and the admin node.

Values = any positive integer number.

Accepted by:

- Cluster definition file
- `cattr` command
- `cm node set`
- `discover` command

VLAN and general network options

`cmcs_per_mgmt_vlan` (HPE Apollo 9000 clusters only)

Specifies the number of chassis management controllers (CMCs) included in an automatically generated management subnet/VLAN before creating one.

To disable this feature, set its value to `0` before any management switches are discovered on the cluster. When this feature is disabled, all HPE Apollo 9000 non-ICE compute nodes are placed on the `head` and `head-bmc` networks.

Values = must be a multiple of `4`. For example, `4`, `8`, `16`.

Range = `0` <= *arg* <= `48`.

Accepted by:

- Cluster configuration tool
- cadmin command
- cattr command

head_vlan (HPE SGI 8600 clusters)

Specifies the number of the head network VLAN. Hewlett Packard Enterprise recommends that you do not change this value.

Default = 1.

Range = 1 <= *arg* <= 4096.

Accepted by:

- Cluster definition file
- cattr command

mcell_network (HPE SGI 8600 clusters)

Specifies whether the cluster includes liquid cooling cells. This value must be set to yes when the cluster includes liquid cooling cell equipment. This value can be set to yes or no for clusters that do not include liquid cooling cells.

Values = yes (default) or no.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- cadmin command
- cattr command

mcell_vlan (HPE SGI 8600 clusters)

Specifies the liquid cooling cell network VLAN. Hewlett Packard Enterprise recommends that you do not change this value.

Default = 3.

Range = 1 <= *arg* <= 4096.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- cattr command

mgmt_vlan_end

Specifies the last non-ICE compute node rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2999`.

Range = `2` <= *arg* <= `4095`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

`mgmt_vlan_start`

Specifies the first non-ICE compute node rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2001`.

Range = `2` <= *arg* <= `4095`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

`rack_vlan_end` (HPE SGI 8600 clusters)

Specifies the last rack leader VLAN on a cluster. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `1100`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

`rack_vlan_start` (HPE SGI 8600 clusters)

Specifies the first rack VLAN. Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `101`.

Range = `1` <= *arg* <= `4096`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command

`redundant_mgmt_network` (HPE SGI 8600 clusters)

Specifies the default setting for the redundant management network. If no value is supplied to the `discover` command at configuration time, the installer populates all nodes with this attribute value.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

`switch_mgmt_network` (HPE SGI 8600 clusters)

Specifies the default setting for the switch management network. If no value is supplied to the `discover` command at configuration time, the installer populates all nodes with this attribute value.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

Console server options

The admin node and the leader nodes manage other nodes. On a cluster without leader nodes, the admin node manages all the (non-ICE) compute nodes. On a cluster with leader nodes, the admin node manages leader nodes and non-ICE compute nodes. ICE leader nodes manage the ICE compute nodes.

On the management nodes, there are files in the `/var/log/consoles` directory for each node that the admin node or the leader node manages. The files contain log information from the baseboard management controllers (BMCs) on the subordinate nodes. That is, on the admin node, the `/var/log/consoles` directory contains log information for each node under admin node control. On each leader node, the `/var/log/consoles` directory contains log information for each node that reports to the leader node.

The console server options let you control the quantity and frequency of log information that is collected. The cluster manager software logs BMC output to the /var/log/consoles directory. In the /var/log/consoles directory, there is a file for each node in the cluster. If you tune the console server options, you can limit the amount of traffic between the console and the cluster. Set these options if you want to minimize network contention.

conserver_logging

Specifies console server logging. If set to yes, the console server logs messages to the console through IPMItool. This feature uses some network bandwidth.

Values = yes (default) or no.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- cm node set command
- discover command

conserver_ondemand

Specifies console server logging frequency. When set to no, logging is enabled all the time. When set to yes, logging is enabled only when someone is connected.

Values = yes or no (default).

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- cm node set command
- discover command

console_device

Specifies the console device.

Values = the device hostname

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command

- `cm node show` command
- `discover` command

Networking options

mgmt_net_interfaces

Configures the system to associate and set up the specified interface or interface in Linux. By default, `eth0` and `eth1` are used on leader nodes and on non-ICE compute nodes. If you specify more than one, include the addresses in a comma-separated string, and enclose the string in quotation marks (`"  "`). If you use quotation marks on a command line, remember that quotation marks must be escaped with backslash (`\`) characters. If using predictable network names, the specified names are used.

Values = the interface hostname or hostnames

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

mgmt_net_macs

Specifies MAC addresses for the management network. If you specify more than one, include the addresses in a comma-separated string, and enclose the string in quotation marks (`"  "`). If you use quotation marks on a command line, remember that quotation marks must be escaped with backslash (`\`) characters. Specify to avoid network sniffing discovery.

Values = the interface MAC address or MAC addresses

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

mgmt_net_name

Specifies the name of the management network. Used primarily on large clusters without leader nodes that have multiple routed networks for management.

Values = head or a network name

Default = head

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

net

For external InfiniBand switches. Specifies the name of the served management network when discovering a management leaf switch that is dedicated to the supplied management networks.

Values = ib0 or ib1

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- discover command
- cm node set command

Monitoring options

monitoring_ganglia_enabled

Specifies whether monitoring, through Ganglia, is enabled in the cluster. When set to yes, monitoring is enabled.

Values = yes or no (default).

Accepted by:

- Cluster definition file
- cattr command
- cm monitoring ganglia command

monitoring_kafka_elk_alerta_enabled

Specifies whether monitoring, through Kafka, Elasticsearch, and Alerta are enabled in the cluster. When set to `yes`, monitoring is enabled.

Values = `yes` or `no` (default).

Accepted by:

- Cluster definition file
- `cattr` command
- `cm monitoring kafka` command
- `cm monitoring elk` command
- `cm monitoring alerta` command

monitoring_nagios_enabled

Specifies whether monitoring, through Nagios, is enabled in the cluster. When set to yes, monitoring is enabled.

Values = yes or no (default).

Accepted by:

- Cluster definition file
- `cattr` command
- `cm monitoring nagios` command

monitoring_native_enabled

Specifies whether the cluster manager native monitoring is enabled in the cluster. When set to `yes`, monitoring is enabled.

Values = `yes` or `no` (default).

Accepted by:

- Cluster definition file
- `cattr` command
- `cm monitoring native` command

Miscellaneous options

architecture

Specifies the processor architecture type on a node.

Values = `x86_64` or `aarch64`

Default = `x86_64`

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

blademond_scan_interval (HPE SGI 8600 clusters)

Specifies how often the blade monitor detects changes in the blades. For example, you can change blades or do other blade maintenance. In these situations, you could set this option to a high value so the daemon does not run during the maintenance period.

Specifies the sleep time for the blademond daemon. The daemon waits the specified number of seconds in between checking if the CMC slot maps have changed.

Default = 120.

Values = can be 0 or any positive integer.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command

card_type

Specifies the type of management card in the node.

Values = The cluster manager supports the card types defined in the following file:

/opt/clmgr/etc/cmuserver.conf

In the cmuserver.conf file, see the CMU_VALID_HARDWARE_TYPES field.

Example card types are ipmi, ilo, and ilocm.

Default = ipmi

Range = NA

Accepted by:

- Cluster definition file
- cm node set command
- cm node show command
- discover command

cluster_domain

Specifies the cluster domain name. Hewlett Packard Enterprise recommends that users change this value.

Values = must be a standard domain name.

Accepted by:

- Cluster definition file
- Cluster configuration tool
- `cadmin` command
- `cattr`

dhcp_bootfile

Specifies whether to load iPXE or GRUB2 first when a node is first configured. By default, `dhcp_bootfile=grub2`, which means that GRUB2 loads first. If you specify `dhcp_bootfile=ipxe`, however, the server boot agent loads iPXE instead of GRUB2, and then iPXE loads GRUB2.

In some cases, a node can fail to boot over the network with the default settings. For example, a node might hang when it tries to load the kernel and `initrd` during the boot from its system disk. In this case, modify the DHCP boot file setting to load iPXE first, and then have iPXE load GRUB2.

Only certain nodes might require the `dhcp_bootfile=ipxe` specification. You can use the `discover` command or the `cadmin` command to specify the new boot order, as follows:

- To use the `discover` command and configure `n0` with an iPXE boot first, include the following on the command line:

  ```
  # discover --node 0,dhcp_bootfile=ipxe other_options
  ```

  For *other_options*, specify any additional options you need for the configuration. For example, you can specify a cluster definition file that contains many other specifications. The `n 0,dhcp_bootfile=ipxe` argument on this `discover` command line overrides the default boot option in the cluster definition file.

- To use the `cadmin` command and specify that iPXE boot first on `n0`, include the following on the command line:

  ```
  # cm node set --node n0 --dhcp-bootfile ipxe
  ```

To display the boot file specification, use the `cadmin` command. For example:

```
# cadmin --show-dhcp-bootfile --node n0
grub2
```

The DHCP log file messages reside in the following file on the admin node:

```
/var/log/dhcpd
```

Values = `grub2` (default) or `ipxe`.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `discover` command

discover_skip_switchconfig

Signals the installer to omit the switch configuration steps. When set to yes, the installer does not configure the switches. Set this option to yes when you want to perform a quick configuration change, but you do not need to update the switch configuration. This value is not saved in the cluster definition file, but it can be specified there.

Values = yes or no (default).

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- discover command

disk_bootloader

After installation, specifies whether the node can boot from the on-disk bootloader. When enabled, it is no longer possible to control kernel boot parameters centrally.

Values = yes or no

Default = no

Range = NA

Accepted by:

- Cluster definition file
- cadmin command
- cm node set command
- cm node show command
- discover command

domain_search_path

Specifies the domain search path for the cluster.

Values = one or more domains. If you specify multiple domains, use a comma (,) to separate each domain.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- discover command

hostname1

Specifies a site-specific, custom hostname for a node. Users can specify this name when they want to log into the node. The hostname appears in most cluster manager output.

Values = a hostname

Default =

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

`ice`

Used in the cluster definition file for management switches. Specifies whether there are ICE leader nodes or chassis management controllers connected to the switch.

Values = `yes` or `no`

Default = This attribute is required.

Range = NA

Accepted by:

Cluster definition file

`internal_name`

Defines the function of a component in the cluster definition file. Formerly `tempo_name` (now deprecated). This name can match the hostname. This name never changes for the life of the node.

Values = A name such as `service0` or `mgmtsw0`.

Default = NA

Range = NA

Accepted by:

Cluster definition file

`max_rack_irus` (HPE SGI 8600 clusters)

Specifies the maximum number of chassis in the cluster. When you use the `discover` command, the installer autopopulates the database with this value during the rack configuration. This value is not saved in the cluster definition file, but it can be specified there.

Default = `8`.

Range = `1` <= *arg* <= `16`.

Accepted by:

- Cluster definition file
- Cluster configuration tool
- cadmin command
- cattr command
- discover command

name

In the [templates] section of the cluster definition file, the name= field defines the name for a particular node template. For example, specify name=su-leader to define the list of configuration parameters for scalable unit (SU) leader nodes.

Values = a custom name for the node template

Default = NA

Range = NA

Accepted by:

Cluster definition file

predictable_net_names

Specifies whether the cluster uses predictable network names to describe the network interface cards (NICs). When set to yes, predictable network names are enabled.

Values = yes (default) or no.

Accepted by:

- Cluster definition file
- cadmin command
- cattr command
- cm node set command
- discover command

su_leader

When the cluster includes scalable unit (SU) leaders, this attribute specifies the SU leader to which a non-ICE compute node is attached.

Values = the IP address of an SU leader node

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- cadmin command

- `cm node set` command
- `cm node show` command
- `discover` command

`template_name`

Identifies the custom template for the cluster manager to use when configuring this node. The custom template is defined in the cluster definition file.

Values = the name of a template in the cluster definition file

Default = NA

Range = NA

Accepted by:

Cluster definition file

`type`

Specifies the type of external InfiniBand switch or management switches. If not specified for a management switch, the cluster manager uses link layer discovery protocol (LLDP) to determine which switch is connected directly to the admin node.

Values = `leaf` or `spine`

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `discover` command

Configuring InfiniBand fabric software manually

This appendix includes the following topics:

- **About configuring the InfiniBand fabric**
- **Configuring InfiniBand fabric manually**

About configuring the InfiniBand fabric

Hewlett Packard Enterprise strongly recommends that you use the automated tools provided in the cluster manager to configure the InfiniBand fabric software. To configure the fabric using these preferred tools, see the following:

Configuring the InfiniBand subnets

The recommended configuration method uses the InfiniBand management tool, which uses a GUI. To start the tool, enter the following command from the admin node:

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

For general information about the tool, see the following:

HPE Performance Cluster Manager Administration Guide

If you want to configure InfiniBand fabric software and you cannot use the GUI tool, complete the following procedure:

Configuring InfiniBand fabric manually

Configuring InfiniBand fabric manually

The following topics explain how to use the `sgifmcli` command to configure the InfiniBand fabric:

- **Configuring a master fabric**
- **Enabling the InfiniBand fabric failover mechanism**
- **(Conditional) Configuring the InfiniBand fat-tree network topology**

Configuring a master fabric

When configuring the subnet manager master, the following rules apply:

- Log into the admin node to run the `sgifmcli` commands.
- Each InfiniBand fabric must have a subnet manager master.
- There can be at most one subnet manager master per InfiniBand fabric.
- Fabric configuration and administration can be done only through the subnet manager master.
- Fabric configuration becomes active after (re)starting the subnet manager master.
- If there is a standby, the action of deleting a subnet manager master automatically deletes the standby.

Procedure

1. Use the `sgifmcli` command to configure a subnet manager master.

The format for the sgifmcli command to configure a subnet manager master is as follows:

```
sgifmcli --mastersm --init --id identifier --hostname hostname --fabric fabric --topology topology
```

The command variables are as follows:

- For *identifier*, specify any arbitrary string. The --id option creates a master with the name you supply.
- For *hostname*, specify the host from which you want the subnet manager master to launch.
- For *fabric*, specify either ib0 or ib1.
- For *topology*, specify hypercube, enhanced-hypercube, ftree, or balanced-ftree.

For example, on a cluster with ICE leader nodes, the following command configures a master for fabric ib0 on a hypercube cluster:

```
# sgifmcli --mastersm --init --id master_ib0 --hostname r1lead \
--fabric ib0 --topology hypercube
```

2. Repeat the preceding step for each fabric you want to create.
3. Proceed to the following:

   Enabling the InfiniBand fabric failover mechanism

Enabling the InfiniBand fabric failover mechanism

Each subnet manager needs a failover mechanism. If the master subnet manager fails, the standby subnet manager takes over operation of the fabric. The opensm software performs this failover operation automatically. Typically, rack1 is the MASTER for the ib0 fabric and rack2 has the MASTER for the ib1 fabric.

When you enable the InfiniBand failover mechanism, observe the following rules:

- As an option, each InfiniBand fabric can have exactly one standby.
- If a master subnet manager exists, you can create a standby subnet manager.
- When adding a standby after a master has already been defined and started, stop the master and then use the --init option to define the standby. After you define the standby, restart the master.
- A subnet manager master and subnet manager standby for a particular fabric cannot coexist on the same node.

The following procedure describes how to set up the failover mechanism.

Procedure

1. Stop any subnet manager masters that are defined and running.

   For example, use the following command:

```
# sgifmcli --stop --id master_ib0
```

2. Define the subnet manager standby.

   For example:

```
# sgifmcli --standbysm --init --id standby_ib0 \
--hostname r2lead --fabric ib0
```

3. Start the subnet manager master.

For example:

```
# sgifmcli --start --id master_ib0
```

This command automatically starts the subnet manager master and the subnet manager standby for `ib0`.

4. Check the status of the subnet manager.

   For example, to check the status of `ib0`, enter the following:

   ```
   # sgifmcli --status --id master_ib0

   Master SM
   Host = r1lead
   Guid = 0x0008f10403987da9
   Fabric = ib0
   Toplogy = hypercube
   Routing Engine = dor
   OpenSM = running
   Standby SM
   Host = r2lead
   Guid = 0x0008f10403987d25
   Fabric = ib0
   OpenSM = running
   ```

5. Proceed to the following if the cluster has an InfiniBand fat-tree network topology:

   (Conditional) Configuring the InfiniBand fat-tree network topology

(Conditional) Configuring the InfiniBand fat-tree network topology

Complete the procedure in this topic if your cluster has an InfiniBand fat-tree network topology. After the cluster is provisioned, if you add an external switch to the cluster with fat-tree topology, perform this procedure to configure the external InfiniBand switch.

The `discover` command configures external InfiniBand switches. After you run the `discover` command, you can use the `sgifmcli` command to add and initialize an external switch on the InfiniBand system.

The fat-tree topology involves external InfiniBand switches. For the list of supported external switches, see the `sgifmcli` manpage.

InfiniBand switches are of two types: leaf switches and spine switches. Leaf switches connect to ICE compute nodes. Spine switches connect leaf switches together. The integrated InfiniBand switches in cluster systems are considered to be leaf switches. The external InfiniBand switches used to connect the leaf switches together in a fat-tree topology are considered to be spine switches.

The `sgifmcli` command lets you specify the following keywords for fat-tree topologies: `ftree` and `balanced-ftree`. The `balanced-ftree` keyword configures balanced fat-tree. If the fat-tree topology is not balanced, choose `ftree`. If the fat-tree topology is balanced, choose `bftree`.

The `discover --switch` command is equivalent to `sgifmcli --init` and `sgifmcli --add` when adding an external switch. If the external switch is configured not as an external switch, but as a general node, run the `sgifmcli --init` and `sgifmcli --add` commands.

Procedure

1. Verify the following:

- The switch has been configured, with the `discover` command, on the cluster. The switch needs an IP address on the management network.
- The switch is properly connected to the InfiniBand network.
- The admin port of the switch is properly connected to the Ethernet network.

2. Power on the switch.

   For more information, see your switch documentation.

3. From the admin node, use the `sgifmcli` command to initialize the switch.

   The syntax is as follows:

   ```
   sgifmcli --init --ibswitch --model modelname  --id identifier --switchtype [leaf | spine]
   ```

   For example:

   ```
   # sgifmcli --init --ibswitch --model voltaire-isr-2004  --id isr2004 \
   --switchtype spine
   ```

   The preceding example command configures a Voltaire switch ISR2004 with hostname `isr2004` as a spine switch. `isr2004` refers to the admin port of the switch. This procedure assumes that the switch has been configured previously with the `discover` command. The switch is now initialized and the root globally unique identifier (GUID) from the spine switches has been downloaded.

4. From the admin node, use the `sgifmcli` command to add the switch to the fabric.

   The syntax is as follows:

   ```
   sgifmcli --add --id fabric --switch hostname
   ```

   For example, the following command connects `isr2004` is connected to the `ib0` fabric:

   ```
   # sgifmcli --add --id ib0 --switch isr2004
   ```

5. (Conditional) Stop and restart the subnet manager master.

   Complete this step if the subnet master manager was running when you added the switch.

   For example:

   ```
   # sgifmcli --stop --id master_ib0
   # sgifmcli --start --id master_ib0
   ```

6. Restart the subnet manager master and the optional subnet manager standby.

   For example:

   ```
   # sgifmcli --start --id master_ib0
   ```

   If you define a standby, the standby assumes control over the switch if the subnet manager master fails.

Predictable network interface card (NIC) names

The following topics contain information about predictable NIC names:

- **About predictable network interface card (NIC) names**
- **Retrieving NIC names**
- **Disabling predictable names**

About predictable network interface card (NIC) names

By default, the cluster manager assigns **predictable names** to the Ethernet NICs within a node. This practice ensures that each NIC name is boot persistent. Predictable names are different for different types of nodes with different types of motherboards. Predictable names are the same across like hardware. For example, if your cluster has only one type of compute node, then the predictable names are the same for all compute nodes in the cluster. When an ICE leader node uses predictable names, all ICE compute nodes under that ICE leader node use predictable names.

The cluster manager also supports legacy names as NIC names. For example, `eth0`, `eth1` are legacy names. Legacy NIC names can change when you boot the cluster. For example, assume that the cluster includes multiple adapters and NICs in a given node. For this cluster, the Linux mechanisms that maintain persistent names in the wanted order can fail to rename NICs properly.

NOTE: Do not mix predictable NIC names and legacy NIC names in the same cluster. HPE does not use predictable names for InfiniBand devices.

The following table shows comparable predictable NIC names and legacy NIC names for an example cluster.

Table 10: Example cluster - using predictable NIC names and legacy NIC names

Node type and role	Network role	Example predictable name	Example legacy name
CH-C1104-GP2 admin node	House network	`ens20f0`	`eth0`
	Management #1	`ens20f1`	`eth1`
	Management #2	`ens20f2`	`eth2`
CH-C1104-GP2 leader node	Management #1	`ens20f0`	`eth0`
	Management #2	`ens20f1`	`eth1`
ICE-XAIP129 ICE compute	Management network	`enp6s0`	`eth0`
C1104-TY13 non-ICE compute	Management network	`enp1s0f0`	`eth0`
	House network	`enp1s0f1`	`eth1`

Retrieving NIC names

The tools you can use to retrieve predictable NIC names differ depending on your situation. The following topics show how to retrieve NIC names for different situations:

- **Retrieving predictable NIC names by running the `configure-cluster` command on the admin node**
- **Retrieving predictable NIC names by examining the cluster definition file**
- **Retrieving predictable NIC names by logging into a node**

Retrieving predictable NIC names by running the `configure-cluster` command on the admin node

You can use the `configure-cluster` command to retrieve predictable network names for the network interfaces on admin nodes. This method returns information for the following:

- The administrative node of a highly available (HA) cluster system. This node is actually a virtual machine (VM).
- The physical admin node of a non-HA cluster system.

The following procedure explains how to retrieve the following:

- The predictable network interface name for the house interface
- The predictable network name for the management network interfaces

Procedure

1. Log into the administrative node as the root user.

   On HA clusters, this node is the VM admin node. On non-HA clusters, this node is the only admin node, which is a physical admin node.

2. At the command prompt, enter the following:

   ```
   # configure-cluster
   ```

3. Note the message in the popup window that the cluster manager displays.

   For example, the message might be the following:

   ```
   Used cached values for house network (ens20f0) and/or Management
   Network Interfaces (ens20f1,ens20f2). Remove
   /etc/opt/sgi/configure-cluster-ethernets to be prompted again or
   use the cluster definition file.
   ```

   This message indicates that the house interface for the admin node is `ens20f0`.

Retrieving predictable NIC names by examining the cluster definition file

The cluster definition file resides on the admin node. This file includes system data about nodes, interfaces, MAC addresses, and other cluster characteristics. The cluster definition file can reside anywhere on the admin node, but by default, HPE writes the file to the following location:

```
/var/tmp/mfgconfigfile
```

You can use the following command to write the cluster definition file to a file of your choosing:

```
# discover --show-configfile > filename
```

The installer associates nodes with templates only when the `discover` command runs at installation time. After installation, the cluster manager does not save the node associations with the templates.

The following example shows a cluster definition file:

```
[templates]
name=leader, predictable_net_names=yes, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head,
mgmt_net_bonding_master=bond0, transport=udpcast, redundant_mgmt_network=yes, switch_mgmt_network=yes,
dhcp_bootfile=grub2, conserver_logging=yes, conserver_ondemand=no, tpm_boot=no, disk_bootloader=no,
mgmtsw=mgmtsw0, console_device=ttyS1, mgmt_net_bonding_mode=802.3ad, rootfs=disk,
mgmt_net_interfaces="enp1s0f0,enp1s0f1"

name=ice-compute, rootfs=nfs, console_device=ttyS1, transport=udpcast, bmc_username=ADMIN, bmc_password=ADMIN,
baud_rate=115200, mgmt_net_interfaces="enp7s0"

name=admin, console_device=ttyS1, rootfs=disk

name=compute-arcadia, predictable_net_names=yes, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head,
mgmt_net_bonding_master=bond0, transport=udpcast, redundant_mgmt_network=yes, switch_mgmt_network=yes,
dhcp_bootfile=grub2, conserver_logging=yes, conserver_ondemand=no, tpm_boot=no, disk_bootloader=no,
mgmtsw=mgmtsw0, console_device=ttyS1, mgmt_net_bonding_mode=active-backup, rootfs=disk,
mgmt_net_interfaces="enp1s0f0,enp1s0f1"

name=compute-tripoli, predictable_net_names=yes, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head,
mgmt_net_bonding_master=bond0, transport=udpcast, redundant_mgmt_network=yes, switch_mgmt_network=yes,
dhcp_bootfile=grub2, conserver_logging=yes, conserver_ondemand=no, tpm_boot=no, disk_bootloader=no,
mgmtsw=mgmtsw0, console_device=ttyS1, mgmt_net_bonding_mode=active-backup, rootfs=disk,
mgmt_net_interfaces="enp3s0f0,enp3s0f1", console=ttys0


[nic_templates]
template=leader, network=head-bmc, net_ifs="bmc0"
template=leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="enp1s0f0,enp1s0f1"
template=leader, network=ib-0, net_ifs="ib0"
template=leader, network=ib-1, net_ifs="ib1"

template=ice-compute, network=ib-0, net_ifs="ib0"
template=ice-compute, network=ib-1, net_ifs="ib1"
template=ice-compute, network=gbe, net_ifs="enp7s0"
template=ice-compute, network=bmc, net_ifs="bmc0"

template=compute-arcadia, network=ib-1, net_ifs="ib1"
template=compute-arcadia, network=head, bonding_master=bond0, bonding_mode=active-backup,
net_ifs="enp1s0f0,enp1s0f1"
template=compute-arcadia, network=ib-0, net_ifs="ib0"
template=compute-arcadia, network=head-bmc, net_ifs="bmc0"

template=compute-tripoli, network=ib-1, net_ifs="ib1"
template=compute-tripoli, network=head, bonding_master=bond0, bonding_mode=active-backup,
net_ifs="enp1s0f0,enp1s0f1"
template=compute-tripoli, network=ib-0, net_ifs="ib0"
template=compute-triploi, network=head-bmc, net_ifs="bmc0"

[discover]
internal_name=admin, admin_house_interface=enp1s0f0, mgmt_net_interfaces="enp1s0f1,enp1s0f2"

internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="02:04:96:98:fa:54",
redundant_mgmt_network=yes, net=head/head-bmc, ice=yes, type=spine

internal_name=r1lead, mgmt_bmc_net_macs="00:25:90:ff:63:76",
mgmt_net_macs="00:25:90:fc:96:ec,00:25:90:fc:96:ed",
template_name=leader, ice_template_name=ice-compute

internal_name=compute-arcadia, mgmt_bmc_net_macs="00:25:90:ff:61:b1",
mgmt_net_macs="00:25:90:fc:96:e4,00:25:90:fc:96:e5", template_name=compute-arcadia

internal_name=compute-tripoli, mgmt_bmc_net_macs="00:25:90:ff:61:b2",
mgmt_net_macs="00:25:90:fc:96:e4,00:25:90:fc:96:e6", template_name=compute-tripoli

[dns]
cluster_domain=smc-default.americas.hpe.com
nameserver1=137.38.31.248
nameserver2=137.38.224.40
nameserver3=137.38.225.5

[attributes]
rack_vlan_start=101
udpcast_max_wait=10
conserver_logging=yes
mcell_vlan=3
max_rack_irus=16
rack_vlan_end=1100
udpcast_min_wait=10
```

```
redundant_mgmt_network=yes
udpcast_rexmit_hello_interval=0
domain_search_path=ib0.smc-default.americas.com,americas.hpe.com,engr.hpe.com,corp.hpe.com
conserver_ondemand=no
udpcast_mcast_rdv_addr=224.0.0.1
udpcast_max_bitrate=900m
udpcast_min_receivers=1
mcell_network=yes
head_vlan=1
dhcp_bootfile=grub2
blademond_scan_interval=120
switch_mgmt_network=yes

[networks]
name=public, subnet=137.38.82.0, netmask=255.255.255.0, gateway=137.38.82.254
name=head, type=mgmt, vlan=1, subnet=172.23.0.0, netmask=255.255.0.0, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0, netmask=255.255.0.0
name=mcell-net, type=cooling, subnet=172.26.0.0, netmask=255.255.0.0
name=ha-net, type=ha, subnet=192.168.161.0, netmask=255.255.255.0
name=ib-0, type=ib, subnet=10.148.0.0, netmask=255.255.0.0
name=ib-1, type=ib, subnet=10.149.0.0, netmask=255.255.0.0
name=gbe, type=lead-mgmt, subnet=10.159.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
name=bmc, type=lead-bmc, subnet=10.160.0.0, netmask=255.255.0.0, rack_netmask=255.255.252.0
```

In the preceding file, the information that pertains to predictable network interface names appears in **bold print**. The file includes a `[templates]` section. When you use a `[templates]` section, you avoid the need to include a `mgmt_net_interfaces=`*`id`* field for each node in the cluster. For example, assume that the cluster has 100 identical non-ICE compute nodes in a cluster. For this cluster, a single entry in the `[templates]` section can provide the cluster manager with all the information it needs for all 100 nodes. Without a `[templates]` section, the cluster configuration file requires 100 entries for each of the 100 non-ICE compute nodes.

Retrieving predictable NIC names by logging into a node

The cluster manager attempts to bring up the management network even when the predictable NIC names are missing or incorrect. No bonding is set up for this debugging situation, and the cluster manager issues a warning message as the node boots up. If the boot fails, you can access the serial console to diagnose.

Assuming that the management network is up, you can use the `ip addr show` command to retrieve the predictable NIC name for any node. In the command output, search for the interface that is up.

For example:

```
# ssh r1lead
Last login: ...
.
.
.
# ip addr show
1: lo:  mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
10: ens20f0:  mtu 1500 qdisc mq master bond0 state UP group default qlen 1000
    link/ether 00:25:90:fd:3d:74 brd ff:ff:ff:ff:ff:ff
11: ens20f1:  mtu 1500 qdisc mq master bond0 state UP group default qlen 1000
    link/ether 00:25:90:fd:3d:74 brd ff:ff:ff:ff:ff:ff
12: ens20f2:  mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether 00:25:90:fd:3d:76 brd ff:ff:ff:ff:ff:ff
13: ens20f3:  mtu 1500 qdisc mq state UP group default qlen 1000
    link/ether 00:25:90:fd:3d:77 brd ff:ff:ff:ff:ff:ff
    inet 192.168.161.1/24 brd 192.168.161.255 scope global ens20f3
       valid_lft forever preferred_lft forever
    inet6 fe80::225:90ff:fefd:3d77/64 scope link
       valid_lft forever preferred_lft forever
```

```
14: bond0:  mtu 1500 qdisc noqueue state UP group default qlen 1000
    link/ether 00:25:90:fd:3d:74 brd ff:ff:ff:ff:ff:ff
    inet 172.23.0.3/16 brd 172.23.255.255 scope global bond0
       valid_lft forever preferred_lft forever
    inet 172.24.0.3/16 brd 172.24.255.255 scope global bond0:bmc
       valid_lft forever preferred_lft forever
    inet 172.23.0.2/16 brd 172.23.255.255 scope global secondary bond0:1
       valid_lft forever preferred_lft forever
    inet6 fe80::225:90ff:fefd:3d74/64 scope link
       valid_lft forever preferred_lft forever
15: vlan101@bond0:  mtu 1500 qdisc noqueue state UP group default qlen 1000
    link/ether 00:25:90:fd:3d:74 brd ff:ff:ff:ff:ff:ff
    inet 10.160.0.2/22 brd 10.160.3.255 scope global vlan101
       valid_lft forever preferred_lft forever
    inet 10.159.0.2/22 brd 10.159.3.255 scope global vlan101:gbe
       valid_lft forever preferred_lft forever
    inet 10.160.0.1/22 brd 10.160.3.255 scope global secondary vlan101:1
       valid_lft forever preferred_lft forever
    inet 10.159.0.1/22 brd 10.159.3.255 scope global secondary vlan101:2
       valid_lft forever preferred_lft forever
    inet6 fe80::225:90ff:fefd:3d74/64 scope link
       valid_lft forever preferred_lft forever
16: ib0:  mtu 65520 qdisc pfifo_fast state DOWN group default qlen 256
    link/infiniband 80:00:02:08:fe:80:00:00:00:00:00:00:e4:1d:2d:03:00:6f:51:e1 brd
        00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff
    inet 10.148.0.2/16 brd 10.148.255.255 scope global ib0
       valid_lft forever preferred_lft forever
17: ib1:  mtu 65520 qdisc pfifo_fast state DOWN group default qlen 256
    link/infiniband 80:00:02:09:fe:80:00:00:00:00:00:00:e4:1d:2d:03:00:6f:51:e2 brd
        00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff
    inet 10.149.0.2/16 brd 10.149.255.255 scope global ib1
       valid_lft forever preferred_lft forever
```

Disabling predictable names

To disable the predictable names feature, use the instructions in the README file on the cluster manager installation DVD.

If the cluster has a highly available admin node, it cannot use predictable NIC names on the physical admin nodes. The physical admin nodes require legacy names.

Managing node additions and deletions on large cluster systems

On cluster systems with thousands of nodes, the `discover` command can take a long time to add or delete nodes from the cluster. As an alternative, you can use the `fastdiscover` command. The `fastdiscover` command completes the task in a smaller amount of time.

Do not use the `fastdiscover` command to add new leader nodes or new management switches into the cluster. The command does not support management switch configuration for VLANs or trunking.

The following procedure explains how to use `fastdiscover` as an alternative to `discover`.

Procedure

1. Log into the admin node as the root user.
2. Use the information in the following topic to help you create a cluster definition file for the new nodes you want to add:

   Cluster definition file contents
3. Enter the following command to add the new nodes to the cluster database:

   ```
   fastdiscover new_config_file
   ```

   For *new_config_file*, specify the name of the cluster definition file you created in the following step:

   Step **2**
4. Update the internal configuration files:

   ```
   # update-configs --node admin --suleaders --leaders
   ```

   The command in this step is safe to run even if the cluster does not have scalable unit (SU) leader nodes.
5. Generate boot files:

   ```
   # cm node refresh netboot -n '*'
   ```

   The command in this step can be entered with the following variations:

   - Rather than include `-n '*'` on the command line, you can specify individual nodes.
   - To omit nodes that are already configured and in the cluster database, include the following parameter on the command line:

     ```
     --skip-existing-nodes
     ```
6. Back up the cluster configuration.

   At this time, continue to the following procedure to back up the cluster configuration files:

   Backing up the cluster

Configuring a new switch

New switches require some preliminary configuration before you run the `discover` command to configure them into a cluster. After you complete the preliminary configuration, you can run the `discover` command from the admin node.

The procedures in this topic apply to both stacked and nonstacked switches. Complete the procedures in this topic under the following circumstances:

- You want to add a switch to the cluster.
- You want to replace an existing switch for which you have no backup data. In this situation, proceed as if you want to add a switch.

The procedures support the following types of switches:

- HPE FlexFabric switches
- HPE FlexNetwork switches
- Extreme Networks switches

NOTE: To replace an existing switch for which you have backup data, use the procedure in the following:

HPE Performance Cluster Manager Administration Guide

The procedures are as follows:

- **Preparing to configure an Extreme Networks switch**
- **Preparing to configure an HPE FlexFabric switch or an HPE FlexNetwork switch**
- **Running the `discover` command for a new switch**

Preparing to configure an Extreme Networks switch

Procedure

1. Access the switch through a console cable.
2. Log in with the default credentials.

   These credentials are one of the following:

   - Username = `admin`, password = `admin`

     Or

   - Username = `admin`, password = <*blank*>

     For <*blank*>, simply press Enter.

3. Enter the following commands:

```
enable dhcp vlan default
enable flooding all_cast ports all
enable jumbo-frame ports all
enable lldp ports all
enable loopback-mode vlan default
```

4. Enter the following command to retrieve the switch MAC address:

   ```
   show switch  | grep MAC
   ```

   For example:

   ```
   Slot-1 mgmtsw8.3 # show switch  | grep MAC
   System MAC:       02:04:96:8B:CC:A8
   ```

   Record the MAC address that this command returns. The cluster definition file requires you to specify the switch MAC address in a slightly different format. To specify the MAC address in this example in the cluster definition file, reformat the address as follows:

   ```
   mgmt_net_macs="02:04:96:8b:cc:a8"
   ```

5. Proceed to the following:

   Running the `discover` command for a new switch

Preparing to configure an HPE FlexFabric switch or an HPE FlexNetwork switch

Procedure

1. Access the switch through a console cable.

2. Log in with the default credentials.

   The username is `admin`, and the password is `admin`.

3. Enter the following commands:

   ```
   system-view
   interface Vlan-interface 1
   ip address dhcp-alloc
   quit
   local-user admin
   password simple admin
   service-type telnet
   authorization-attribute user-role network-admin
   quit
   telnet server enable
   line vty 0 63
   authentication-mode scheme
   user-role network-admin
   quit
   undo stp global enable
   save safely force
   ```

4. Enter the following command to retrieve the switch MAC address:

   ```
   display int vlan 1 | include hardware
   ```

For example:

```
display int vlan 1 | include hardware
IP packet frame type: Ethernet II, hardware address: d894-03fe-07b1
```

Record the MAC address that this command returns. The cluster definition file requires you to specify the switch MAC address in a slightly different format. To specify the MAC address in this example in the cluster definition file, reformat the address as follows:

```
mgmt_net_macs="d8:94:03:fe:07:b1"
```

5. Proceed to the following:

   Running the discover command for a new switch

Running the discover command for a new switch

Procedure

1. Log into the admin node as the `root` user.
2. Edit the cluster definition file to include the new switch.

   Example 1. The following is an example for a spine switch:

   ```
   # spine switch example
   internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="02:04:96:8b:cc:a8",
   redundant_mgmt_network=yes, net=head/head-bmc, type=spine, ice=yes
   ```

   Example 2. The following is an example for a leaf switch:

   ```
   # leaf switch example
   internal_name=mgmtsw1, mgmt_net_name=head,
   mgmt_net_macs="d8:94:03:fe:07:b1", redundant_mgmt_network=yes, net=head/head-bmc, type=leaf, ice=yes
   ```

   It is possible that you cannot locate the cluster definition file. In this case, see the following topic for information about how to create a new one in a location of your choosing:

   Preparing to install the operating system and the cluster manager jointly

   For more information and for cluster definition file examples, see the following:

   Cluster definition file examples with node templates, network interface card (NIC) templates, and predictable names

3. Run the `discover` command in the following format:

   ```
   discover --configfile path_to_CDF --mgmtswitch X
   ```

   For *path_to_CDF*, specify the full path to your cluster definition file. By default, the file is stored in the following location:

   ```
   /var/tmp/mfgconfigfile
   ```

   For *X*, specify the number for the new switch.

4. Enter the following command to determine the firmware version that matches this installation of the cluster manager:

   ```
   switchconfig sanity_check -s mgmtswX | grep firmware
   ```

   For *X*, specify the switch number.

Example 1: Example Extreme Networks command:

```
admin:~ # switchconfig sanity_check -s mgmtsw8 | grep firmware
checking switch firmware on mgmtsw8 ...
        Switch installed in Slot-1 has firmware 16.2.5.4 installed (good)
        Switch installed in Slot-2 has firmware 16.2.5.4 installed (good)
```

Example 2: Example HPE switch command:

```
admin:~ # switchconfig sanity_check -s mgmtsw6 | grep firmware
checking switch firmware on mgmtsw6 ...
        mgmtsw6 slot 1 (5510 24G 4SFP+ HI 1-slot Switch) has firmware '7.1.070
           Release 1309P07-US' installed (recommended: '7.1.070 Release 1309P07' or
           '7.1.070 Release 1309P07-US')
        mgmtsw6 slot 2 (5510 24G 4SFP+ HI 1-slot Switch) has firmware '7.1.070
           Release 1309P07-US' installed (recommended: '7.1.070 Release 1309P07' or
           '7.1.070 Release 1309P07-US')
```

5. Upgrade the switch firmware as needed.

Configuring a cluster that uses an unsupported Ethernet switch

The cluster manager supports the Ethernet switches as described in the cluster manager release notes. An advantage to using supported Ethernet switches is that you can use cluster manager tools, such as `switchconfig`, to manage them.

If the cluster includes switches that are not supported, modify the installation procedure according to the steps in this topic. Use commands specific to that switch to complete some configuration steps manually.

Unsupported switches are included in the cluster as unmanaged switches. For these switches, the cluster manager does not attempt to automatically configure any switch settings.

The following procedure explains how to configure an unsupported switch into a cluster.

Procedure

1. Enter the following command:

   ```
   # cadmin --enable-discover-skip-switchconfig
   ```

   This command accomplishes the following:

   - It prevents the cluster manager from logging into management switches at a global level.
   - It allows you to configure the unsupported switches later in the installation.

2. Configure the switches for multicast or configure the cluster manager to use unicast.

   This step ensures that the leader and compute nodes receive their images from the admin node in an efficient manner. Do one of the following:

   - Verify whether the unsupported switch is configured for **IGMP** and **IGMP Snooping**. Configure those two settings if they are not in effect at this time. The cluster manager uses a multicast protocol called UDPcast to image leader and compute nodes during the boot process. For multicast to be successful, the management switches must support IGMP and IGMP Snooping. For information, see the switch configuration documentation.

     Or

   - Configure the cluster manager to use BitTorrent when it images the compute nodes. BitTorrent is not a multicast method. It is unicast.

     For information about how to change the method by which the leader and compute nodes receive images, see the following:

     Node provisioning takes too long or fails to complete

3. Return to the following procedure and complete the installation, which includes running the `discover` command:

   Completing the admin node software installation

   The `discover` command configures supported switches and all other components to be under cluster manager control. After the installation, consider one of the following:

   - Enabling DHCP on the unsupported switch
   - Configuring a static IP address on the unsupported switch

For information, see the documentation for the unsupported switch. DHCP enables the cluster manager to assign an IP address to the switch. To manage these switches remotely, do the following for the switch:

- Enable either Telnet or SSH.
- Create a remote username and strong password.

Because you ran the `cadmin --enable-discover-skip-switchconfig` command before you ran the `discover` command, the `discover` command allows DHCP to assign supported switches an IP address. In this way, you can SSH or Telnet to the supported switches if necessary. Assigning a static IP achieves the same outcome. That is, the management switch has an entry in `/etc/hosts`, but the cluster manager does not remotely log into the switch automatically.