Hewlett Packard Enterprise

# HPE Performance Cluster Manager Installation Guide for Clusters With Scalable Unit (SU) Leader Nodes

**Abstract**

This publication describes how to install the HPE Performance Cluster Manager 1.6 software on an HPE cluster system with scalable unit (SU) leader nodes.

Part Number: P36611-003
Published: November 2021
Edition: 3

## Notices

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

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

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

## Acknowledgments

Intel®, Itanium®, Optane™, Pentium®, Xeon®, Intel Inside®, and the Intel Inside logo are trademarks of Intel Corporation or its subsidiaries.

AMD and the AMD EPYC™ and combinations thereof are trademarks of Advanced Micro Devices, Inc.

Microsoft® and Windows® are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

Adobe® and Acrobat® are trademarks of Adobe Systems Incorporated.

Java® and Oracle® are registered trademarks of Oracle and/or its affiliates.

UNIX® is a registered trademark of The Open Group.

All third-party marks are property of their respective owners.

## Product-specific acknowledgments

ARM® is a registered trademark of ARM Limited.

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

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

## Revision history

| Part number | Publication date | Edition | Summary of changes |
|---|---|---|---|
| P36611-003 | November 2021 | 3 | Supports the HPE Performance Cluster Manager 1.6 release.<br><br>Edition 3 replaces Edition 2. Edition 3 includes information about the quorum high-availability software installation and includes other corrections and additions. |
| P36611-003 | September 2021 | 2 | Supports the HPE Performance Cluster Manager 1.6 release. The HPE Performance Cluster Manager 1.6 release package includes Edition 1.<br><br>Edition 2 replaces Edition 1 and includes enhancements to the scalable unit (SU) leader node upgrade procedure. |
| P36611-003 | September 2021 | 1 | Supports the HPE Performance Cluster Manager 1.6 release. |
| P36611-002 | March 2021 | 1 | Supports the HPE Performance Cluster Manager 1.5 release. |
| P36611-001 | September 2020 | 1 | Original publication. Supports the HPE Performance Cluster Manager 1.4 release. |

# 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.

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 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.

To start a bare-metal installation, proceed to the following:

**Installing the operating system and the cluster manager simultaneously on the admin node**

## HPE Performance Cluster Manager operating system releases supported

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 tables show the releases that the HPE Performance Cluster Manager 1.6 release supports.

**Table 1: Operating systems for x86_64 architectures**

| Node type | Operating systems supported |
|---|---|
| Admin | RHEL 8.4 |
| | RHEL 7.9 |
| | SLES 15 SP3 |
| | SLES 12 SP5 |
| | CentOS 7.9 |
| Leader | RHEL 8.4 |
| | RHEL 7.9 |
| | SLES 15 SP3 |
| | SLES 12 SP5 |
| | CentOS 7.9 |
| Compute | RHEL 8.4 |
| | RHEL 7.9 |
| | SLES 15 SP3 |
| | SLES 12 SP5 |
| | CentOS 8.4 |
| | CentOS 7.9 |
| | HPE Cray operating system (COS) 2.2 |

**Table 2: Operating systems for Arm (AArch64) architectures**

| Node types | Operating systems supported |
|---|---|
| Admin | SLES 15 SP3 |
| Leader | SLES 15 SP3 |
| Compute | RHEL 8.4 |
| | SLES 15 SP3 |
| | SLES 12 SP5 |

The following information pertains to RHEL 8.X support:

- 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**

On HPE Cray EX systems and on HPE Apollo 9000 systems, the cluster manager supports only SLES 15 SP3 and RHEL 8.4.

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**. Also see the HPE Performance Cluster Manager release notes.

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 scalable unit (SU) leader nodes must be x86_64 servers. For HPE Cray EX clusters, the admin node must be an HPE ProLiant DL325 server. For HPE Apollo 9000 clusters, the admin node must be an HPE ProLiant DL360 server.

You can configure admin nodes for high availability (HA). The cluster manager supports two HA admin node configurations:

- Quorum high availability (quorum HA). Consists of three nodes and a Gluster file system.

  The cluster manager supports RHEL 8.4 and SLES 15 SP3 on quorum HA admin nodes. The cluster manager does not support CentOS on quorum HA admin nodes.

- System admin controller high availability (SAC HA). Consists of two nodes with the specific storage and network configuration dictated in the presales phase. To deploy a SAC HA admin node, purchase and download the HA operating system software offerings from the operating system vendor.

  The following restrictions apply to clusters with SAC HA admin nodes:

  - The cluster manager supports RHEL 8.4, SLES 15 SP3, and SLES 12 SP5 on SAC HA admin nodes.
  - The cluster manager requires SAC HA admin nodes to be x86_64 HPE ProLiant DL360 servers.

The following additional information applies to clusters with SU leader nodes:

- The operating system used on the SU leader nodes must match the operating system used on the admin node.
- The cluster manager supports only x86_64 servers as SU leader nodes.

The following HPE Performance Cluster Manager ISO images are available for each operating system:

- The admin installer ISO, which is bootable and which installs the operating system and cluster manager simultaneously. Hewlett Packard Enterprise recommends that you use this ISO image.
- The repository ISO, which installs the cluster manager onto a pre-installed operating system.

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

# Cluster manager documentation

The following list shows the HPE Performance Cluster Manager documentation:

- The release notes contain feature information, platform requirements, and other release-specific guidance. To access the release notes, follow the links on the following website:

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

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

  `/docs`

  Hewlett Packard Enterprise strongly recommends that you read the release notes, particularly the Known Issues section and the Workarounds section.

- The following guide presents an overview of the cluster manager and explains how to attach a factory-installed cluster to your site network:

  **HPE Performance Cluster Manager Getting Started Guide**

- The bare-metal installation documentation is specific to each platform. These guides are as follows:

  - **HPE Performance Cluster Manager Installation Guide for Clusters With ICE Leader Nodes**
  - **HPE Performance Cluster Manager Installation Guide for Clusters With Scalable Unit (SU) Leader Nodes**
  - **HPE Performance Cluster Manager Installation Guide for Clusters Without Leader Nodes**

- The following guide explains the power management features included in the cluster manager:

  **HPE Performance Cluster Manager Power Management Guide**

- The following guide includes procedures and information about system-wide administration features:

  **HPE Performance Cluster Administration Guide**

- The following quick-start guide presents an overview of the installation process:

  **HPE Performance Cluster Manager Installation Quick Start**

- The following command reference shows the cluster manager commands and compares them with the commands used in the SGI Management Suite and in the HPE Insight Cluster Manager Utility:

  **HPE Performance Cluster Manager Command Reference**

After installation, the documentation reside on the system in the following directories:

- Release notes and user guides: `/opt/clmgr/doc`
- Manpages: `/opt/clmgr/man`

# `cm` command information

Many cluster manager commands are of the following form:

`cm` *`topic [subtopic ...] action parameters`*

The `cm` commands support tab completion for each *topic*, each *subtopic*, each *action*, and many parameters.

The cluster manager implements tab completion for the `-i` *`image`* and the `--image` *`image`* options by comparing command input against the image names stored in the HPE Performance Cluster Manager database.

You can use wildcard characters in the cluster manager `cm` commands. If you use wildcards in the `cm` commands, enclose your specification in apostrophes (`' '`). The following table shows the most commonly used wildcard characters.

**Table 3: Wildcard characters**

| Wildcard | Effect |
|---|---|
| * | Matches one or more characters. For example, the following specifies all nodes in rack 1, chassis 1, tray 1 on an HPE Apollo 9000 cluster:<br>`'r1c1t1n*'` |
| ? | Matches exactly one character. For example, the following specifies all nodes in rack 1 that have a single-character chassis:<br>• On an HPE Apollo 9000 cluster: `'r1c?t*n*'`<br>• On an HPE SGI 8600 cluster: `'r1i?n*'` |
| [ ] | Matches any of the range of characters specified within brackets. For example, the following specifies racks `11`, `12`, `13`, and `14`: `'rack1[1-4]'` |

# Node identification

The cluster manager recognizes distinct node hostnames for each type of cluster that it supports.

**NOTE:** The information in this topic shows the compute node names that the cluster manager assigns to nodes by default. This naming scheme identifies components by their location in the cluster. These names are assigned automatically when the compute nodes are configured into the cluster.

### HPE Cray EX node identification

On HPE Cray EX supercomputers, the node name is in the following format:

`x`*`CABINET`*`c`*`CHASSIS`*`s`*`SLOT`*`b`*`BLADE`*`n`*`NODE`*

The variables are as follows:

| Variable | Specification |
|---|---|
| *CABINET* | A 4-digit cabinet identifier in the range 1 <= *CABINET* <= 9999. Specific cabinet identifiers are as follows:<br>• HPE Cray EX fluid-cooled compute: x1000 - x2999<br>• HPE Cray EX air-cooled I/O: x3000 - x4999<br>• HPE Cray EX air-cooled compute: x5000 - x5999<br>• HPE Cray EX TDS: x9000<br>Examples: `x1004`, `x3001`. |
| *CHASSIS* | A 1-digit chassis identifier in the range 0 <= *CHASSIS* <= 7. Examples: `c1`, `c7`. |

*Table Continued*

| Variable | Specification |
|---|---|
| *SLOT* | A 1-digit slot identifier in the range 0 <= *SLOT* <= 7. Examples: `s1`, `s4`. |
| *BLADE* | A 1-digit blade identifier in the range 0 <= *BLADE* <= 1. Examples: `b0`, `b1`. |
| *NODE* | A 1-digit node identifier in the range 0 <= *NODE* <= 1. Examples: `n0`, `n1`. |

The following are node identification examples:

- `x9000c1s2b0n0` is a compute node.
- `fmn01` and `fmn02` are HPE Slingshot fabric management nodes.

#### HPE Cray EX switch identification

The default switch naming conventions are similar to the default node naming conventions. On HPE Cray EX supercomputers, the switch names are in the following format:

`x`*`CABINET`*`c`*`CHASSIS`*`r`*`SWITCH`*`b`*`BMC`*

The variables are as follows:

| Variable | Specification |
|---|---|
| *CABINET* | A 4-digit rack identifier in the range 1 <= *CABINET* <= 9999. Examples: `x0046`, `x0178`. |
| *CHASSIS* | A 1-digit chassis identifier in the range 1 <= *CHASSIS* <= 4. Examples: `c1`, `c2`. |
| *SWITCH* | A 1-digit tray identifier in the range 0 <= *SWITCH* <= 7. Examples: `r5`, `r7`. |
| *BMC* | A 1-digit switch identifier in the range 0 <= *BMC* <= 1. Examples: `b0`, `b1`. |

For example: `x1203c0r5b0` is a hostname for an HPE Cray EX switch controller.

#### HPE Apollo 9000 node identification

On HPE Apollo 9000 clusters, the node name is in one of the following formats:

`r`*`RACK`*`c`*`CHASSIS`*`t`*`TRAY`*`n`*`NODE`*

The variables are as follows:

| Variable | Specification |
|---|---|
| *RACK* | A 3-digit rack identifier in the range 1 <= *RACK* <= 999. Examples: `r46`, `r178`. |
| *CHASSIS* | A 1-digit chassis identifier in the range 1 <= *CHASSIS* <= 4. Examples: `c1`, `c2`. |

*Table Continued*

| Variable | Specification |
|---|---|
| *TRAY* | A 1-digit tray identifier in the range 1 <= *TRAY* <= 8. Examples: `t5`, `t8`. |
| *NODE* | A 1-digit node identifier in the range 1 <= *NODE* <= 4. Examples: `n1`, `n4`. |

For example: `r100c3t5n1`

#### HPE Apollo 9000 switch identification

The default switch naming conventions are similar to the default node naming conventions. On HPE Apollo 9000 clusters, the switch names are in the following format:

`r`*RACK*`c`*CHASSIS*`t`*TRAY*`s`*SWITCH*

The variables are as follows:

| Variable | Specification |
|---|---|
| *RACK* | A 3-digit rack identifier in the range 1 <= *RACK* <= 999. Examples: `r46`, `r178`. |
| *CHASSIS* | A 1-digit chassis identifier in the range 1 <= *CHASSIS* <= 4. Examples: `c1`, `i2`. |
| *TRAY* | A 1-digit tray identifier in the range 1 <= *TRAY* <= 8. Examples: `t5`, `t8`. |
| *SWITCH* | A 1-digit switch identifier in the range 1 <= *SWITCH* <= 4. Examples: `s2`, `s3`. |

# Installation flow diagram

The following figure summarizes the procedural flow for cluster manager installations on clusters with scalable unit (SU) leader nodes.

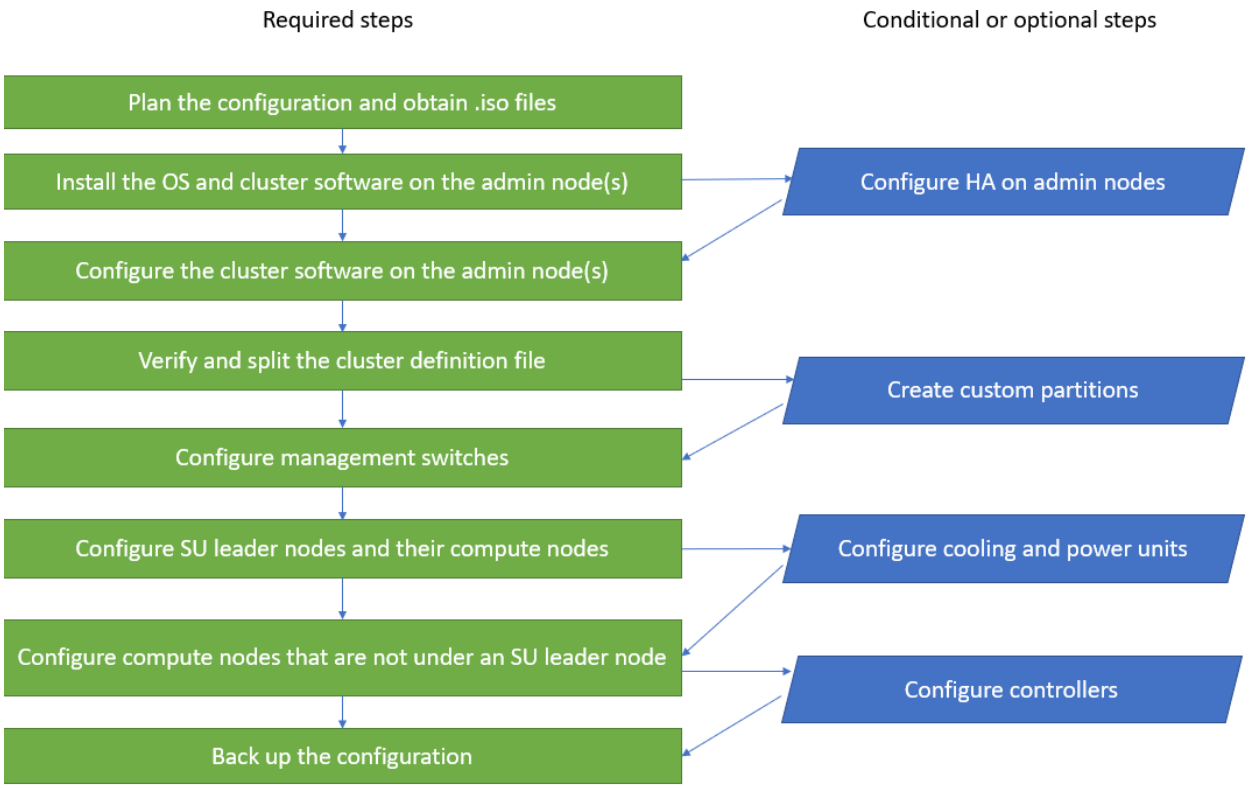


**Figure 1: Installation process flow for clusters with SU leader nodes**

# Installing the operating system and the cluster manager simultaneously on the admin node

**Procedure**

1. **Preparing to install the operating system and the cluster manager simultaneously on the admin node**
2. **(Optional) Configuring custom partitions on the admin node**
3. **Inserting the installation DVD and booting the admin node**
4. Installing an operating system. Use 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. **(Conditional) Configuring the storage unit**
6. **(Conditional) Enabling an input-output memory management unit (IOMMU)**
7. **Verifying the configuration**

**NOTE:** For an alternate way to install the operating system and cluster manager, see one of the following:

- **Installing the operating system and the cluster manager separately**
- **Upgrading the operating system and reinstalling the cluster manager**

## Preparing to install the operating system and the cluster manager simultaneously on the admin node

**Procedure**

1. Confirm that this procedure can work for you by making sure that at least one of the following is true for the cluster:
   - 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 the option 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.
2. Contact your site network administrator to obtain network information for the node controller in the admin node.

   For the admin node controller, obtain the following:

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

Also obtain the IP addresses of the domain name servers (DNSs) on your site network.

**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, or to install a highly available admin node, plan to move a DVD drive from one physical node to another as you configure each one.

   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. Retrieve the updated packages from the HPE customer portal and the operating system vendor.

   You can obtain the cluster manager installation software, including patches and updates, from the following website:

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

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

   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.

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

   Complete this step only if you want a system admin controller highly available (SAC HA) admin node.

   Do not complete this step if you want to configure a quorum HA admin node.

   The SAC 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.

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

6. Attach the cluster to your site network.

   Use the procedure in the following:

   **HPE Performance Cluster Manager Getting Started Guide**

7. Gather information about the cluster components.

   Ideally, obtain the cluster definition file for this cluster. Proceed as follows:

   - If a cluster definition file 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.

     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.

   - If no cluster definition file is available, plan to use the `cm node discover` command to configure nodes into the cluster.

# (Optional) 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. This procedure lets you choose your own layout for the system disk.

**NOTE:** Do not use the custom partitioning feature to specify additional storage. Do not custom partitioning if you need more than one slot.

If you create custom partitions on the admin node, you can create custom partitions on one or more compute nodes.

The partitions on the 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 compute nodes. The following information pertains to custom partitions on compute nodes:

- If the admin node is configured to use default partitions, you can create custom partitions on compute nodes.
- If the admin node is configured to use custom partitions, you can create custom partitions on compute nodes that use a different partitioning scheme.
- You can create custom partitions on any compute node, and the partitions can be different on each compute node. Create one custom partitioning file for each partitioning scheme that you want to impose on one or more compute nodes.

**NOTE:** Custom partitions do not apply to compute nodes configured with an NFS file system or a `tmpfs` file system. In addition, custom partitions do not apply to compute nodes installed by using AutoYaST or Kickstart.

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. The node discovery commands configure the nodes. When you run the node discovery commands, you can create the same (or different) custom partitions on the compute nodes.

You can create custom partitions on leader nodes, but Hewlett Packard Enterprise recommends that you accept the default partitions on leader nodes.

**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.

Custom partitions do not apply to compute nodes configured with an NFS root file system or a `tmpfs` root file system.

The cluster manager does not support custom admin node partitions on clusters with HA admin 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 compute nodes with multiple slots.

4. Decide where you want `custom_partitions_example.cfg` to reside.

   Typically, you 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.

   Alternatively, you can write the configuration file to the installation media, but this requires assistance from Hewlett Packard Enterprise.

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

6. Save and close the file as `custom_partitions.cfg`
7. (Conditional) Repeat the steps in this procedure on the second or third admin node.

   Repeat this procedure on the additional admin nodes that you plan to configure into a high availability (HA) admin node configuration.

# Inserting the installation DVD and booting the admin node

**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:

   **HPE Performance Cluster Manager operating system releases supported**
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 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 on the installation menus.

   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.

     For more information, see the following:

     **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 an SAC-HA or Quorum-HA Physical Host? (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):**

  Supply additional parameters as follows:

  - To configure an HA admin node based on an HPE ProLiant DL360 or an HPE ProLiant DL325, specify a `console=`*`type`* parameter. You need to determine whether to specify `console=ttyS0` or `console=ttyS1` parameter based on the admin node type. For most HPE ProLiant admin nodes, specify `console=ttyS0`.

    For example:

    ```
    console=ttyS0,115200n8
    ```

  - To specify additional boot parameters, enter them in a comma-separated list and press Enter.
  - To configure custom partitions, add the target for the custom partitions file, as follows:

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

    The variables are as follows:

    | 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. |

  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.

   If the system issues a failure message, scroll up to the top of the message to display the steps that explain how to recover the installation. Complete the recovery steps, and continue with this procedure.

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. (Conditional) Repeat the steps in this procedure on the second or third admin node.

    Repeat this procedure on the additional admin nodes that you plan to configure into a high availability (HA) admin node configuration.

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

**Procedure**

1. Use one of the following methods to log into each physical admin node as the root user:

   - Use the intelligent platform management interface (IPMI) tool
   - Use the keyboard, video display terminal, and mouse (KVM) equipment attached to the console or attach your own KVM equipment to the cluster

   To configure a high availability (HA) admin node, complete the steps in this procedure on all physical admin nodes.

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

   ```
   # date
   Fri Apr 20 10:12:50 CDT 2021
   ```

   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:

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

      ```
      # timedatectl list-timezones
      ```

   b. 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 2021-04-15 14:55:33 PDT
      Universal time: Fri 2021-04-15 21:55:33 UTC
      ```

```
RTC time: Fri 2021-04-15 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 2021-03-13 01:59:59 PST
Sun 2021-03-13 03:00:00 PDT
Next DST change: DST ends (the clock jumps one hour backwards) at
Sun 2021-11-06 01:59:59 PDT
Sun 2021-11-06 01:00:00 PST
```

3. 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 fully qualified domain name (FQDN), which is the longer name.

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

4. Complete the following steps to direct network time protocol (NTP) server requests to the server at your site rather than the public time servers of the `pool.ntp.org` project:

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

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

   c. At the end of the file, add lines for the following:

   - Identification for the NTP servers at your site. Add `server xxx.xxx.xxx.xxx iburst` lines.
   - NTP broadcasting to the management network and the baseboard management controller (BMC) network . Add `allow 172.2x.0` lines.

   **NOTE:** Specify the IP addresses of the public time servers at your site. Do not specify the DNS hostnames or FQDNs for the public time servers at your site.

   For example:

   ```
   server 150.166.33.20    iburst
   server 150.166.33.25    iburst
   server 150.166.33.89    iburst
   allow 172.23.0
   allow 172.24.0
   ```

   d. Save and close the file.

5. Use a text editor to open file `/etc/hosts`.

6. For each physical admin node, add an entry in the `/etc/hosts` file that contains the network address and the FQDN for each node.

Use the following format:

```
admin_node_IP admin_node_FQDN admin_node_hostname
```

The variables in the line are as follows:

| Variable | Specification |
|---|---|
| *admin_node_IP1* | The IP address of a physical admin node. |
| [*admin_node_IP2*] | For a single-node admin node, specify the IP address. |
| [*admin_node_IP3*] | For system admin controller high availability (SAC HA) admin nodes, create an additional line for the second admin node.<br><br>For quorum HA admin nodes, create additional lines for the second admin node and the third admin node. |
| *admin_node_FQDN* | The FQDN of the physical admin node. |
| *admin_node_hostname* | The hostname of the physical admin node. |

Example 1. If you have one physical admin node, add a line similar to the following:

```
# physical node address
100.162.244.251 acme-admin.acme.usa.com acme-admin
```

Example 2. If you have two physical admin nodes as part of a SAC-HA configuration, add lines similar to the following:

```
# physical node addresses
   100.162.244.251   acme-admin1.acme.usa.com    acme-admin1
   100.162.244.252   acme-admin2.acme.usa.com    acme-admin2
```

Example 3. If you have three physical admin nodes as part of a quorum HA configuration, add lines similar to the following:

```
# physical node addresses
   100.162.244.251   acme-admin1.acme.usa.com    acme-admin1
   100.162.244.252   acme-admin2.acme.usa.com    acme-admin2
   100.162.244.253   acme-admin3.acme.usa.com    acme-admin3
```

7. Save and close file `/etc/hosts`.
8. Use a text editor to create the following file:

   `/etc/resolv.conf`

   Add the following information to `resolv.conf`, and then save and close the file:

   - The `nameserver` keyword and the IP address of the name server at your site.
   - The `search` keyword and the FQDN.

   For example:

   ```
   search cluster.publicdomain.com publicdomain.com
   nameserver 150.150.39.101
   ```

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

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

10. 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
    .
    .
    .
    ```

11. 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*

12. 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.

---

13. Save and close file `ifcfg-`*`name`*.
14. Reboot the physical admin node, and watch the reboot on the console..

    Wait for the physical admin node to come back up.
15. Log into each physical admin node as the root user.
16. Enter a `ping` command to another server at your site to make sure that the network is functioning.

    Wait for the `ping` to return.
17. Log out of the console window, and use the `ssh` command to log into the admin node as the root user.

# 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.

**Procedure**

1. Use one of the following methods to log into each physical admin node as the root user:

   - Use the intelligent platform management interface (IPMI) tool
   - Use the keyboard, video display terminal, and mouse (KVM) equipment attached to the console or attach your own KVM equipment to the cluster

To configure a high availability (HA) admin node, complete the steps in this procedure on all physical admin nodes.

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

```
# date
Fri Apr 20 10:12:50 CDT 2021
```

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:

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

```
# timedatectl list-timezones
```

**b.** 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 2021-04-15 14:55:33 PDT
Universal time: Fri 2021-04-15 21:55:33 UTC
RTC time: Fri 2021-04-15 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 2021-03-13 01:59:59 PST
Sun 2021-03-13 03:00:00 PDT
Next DST change: DST ends (the clock jumps one hour backwards) at
Sun 2021-11-06 01:59:59 PDT
Sun 2021-11-06 01:00:00 PST
```

3. 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 fully qualified domain name (FQDN), which is the longer name.

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

4. Complete the following steps to direct network time protocol (NTP) server requests to the server at your site rather than the public time servers of the `pool.ntp.org` project:

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

**b.** Search for the following line in `/etc/chrony.conf`:

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

In column 1, replace the exclamation point character (`!`) with a pound sign character (`#`).

**c.** At the end of the file, add lines for the following:

- Identification for the NTP servers at your site. Add `server xxx.xxx.xxx.xxx iburst` lines.
- NTP broadcasting to the management network and the baseboard management controller (BMC) network . Add `allow 172.2x.0` lines.

---

**NOTE:** Specify the IP addresses of the public time servers at your site. Do not specify the DNS hostnames or FQDNs for the public time servers at your site.

---

For example:

```
server 150.166.33.20    iburst
server 150.166.33.25    iburst
server 150.166.33.89    iburst
allow 172.23.0
allow 172.24.0
```

**d.** Save and close the file.

**5.** Enter the following commands to start YaST2:

```
# 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, also enter the following:

```
# export NCURSES_NO_UTF8_ACS=1
```

For information about navigation, see the following:

**YaST navigation**

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

**a.** Select your language.

**b.** Select your keyboard layout.

**c.** Select **Next**.

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

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

**a.** Tab to the box (`[ ] I Agree ...`).

**b.** Press the space bar to accept the license terms. This action puts an `x` in the box, so it looks like this: `[x]`.

**c.** Select **Next**.

**d.** (Conditional) If there are more license agreement screens, select **Next** again.

**9.** On the **Network Settings** screen, prepare to specify the NIC information.

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

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

  **a.** 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
  ```

  **b.** Select **Edit**.

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

  **a.** Highlight the first NIC that appears underneath **Name**.

  **b.** Select **Edit**.

**10.** On the **Network Card Setup** screen, complete the following steps to specify the admin node public NIC:

**a.** Select **Statically Assigned IP Address**. Hewlett Packard Enterprise recommends a static IP address, not DHCP, for the admin node.

**b.** 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.

**c.** In the **Subnet Mask** field, enter the admin node subnet mask.

**d.** In the **Hostname** field, enter the admin node FQDN. HPE requires you to enter an FQDN, not the shorter hostname, into this field. For example, enter `admin.cm.clusterdomain.com`. Failure to supply an FQDN in this field causes the `configure-cluster` command to fail.

**e.** Select **Next**.

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

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

**a.** Select **Hostname/DNS**.

**b.** In the **Hostname** field, enter the admin node hostname.

**c.** (SLES 12 SPX only) In the **Domain Name** field, enter the domain name for your site.

**d.** (SLES 12 SPX only) Put an `X` in the box next to **Assign Hostname to Loopback IP**.

**e.** In the **Name Servers and Domain Search List**, enter the IP addresses of the name servers for your house network.

**f.** In the **Domain Search** field, enter the domains for your site.

**g.** Back at the top of the screen, select **Routing**.

The **Network Settings > Routing** screen appears.

**h.** In the **Default IPV 4 Gateway** field, enter your site default gateway.

**i.** Select **Next**.

**12.** On the **Local User** screen, complete one of the following actions:

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

  or

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

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

**a.** 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.

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

**c.** 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.

**d.** Select **Next**.

**e.** (Conditional) Confirm the password on the popup that appears.

Complete this step if a password popup appears.

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

**15.** Use a text editor to open file `/etc/hosts`.

**16.** For each physical admin node, add an entry in the `/etc/hosts` file that contains the network address and the FQDN for each node.

Use the following format:

*`admin_node_IP admin_node_FQDN admin_node_hostname`*

The variables in the line are as follows:

| Variable | Specification |
|---|---|
| *admin_node_IP* | The IP address of an admin node. |
| [*admin_node_IP2*] | For a single-node admin node, specify the IP address. |
| [*admin_node_IP3*] | For system admin controller high availability (SAC HA) admin nodes, create an additional line for the second admin node.<br><br>For quorum HA admin nodes, create additional lines for the second admin node and the third admin node. |
| *admin_node_FQDN* | The FQDN of the admin node. |
| *admin_node_hostname* | The hostname of the admin node(s). |

Example 1. If you have one physical admin node, add a line similar to the following:

```
# physical node address
100.162.244.251 acme-admin.acme.usa.com acme-admin
```

Example 2. If you have two physical admin nodes as part of a SAC-HA configuration, add lines similar to the following:

```
# physical node addresses
   100.162.244.251   acme-admin1.acme.usa.com    acme-admin1
   100.162.244.252   acme-admin2.acme.usa.com    acme-admin2
```

Example 3. If you have three physical admin nodes as part of a quorum HA configuration, add lines similar to the following:

```
# physical node addresses
   100.162.244.251   acme-admin1.acme.usa.com    acme-admin1
   100.162.244.252   acme-admin2.acme.usa.com    acme-admin2
   100.162.244.253   acme-admin3.acme.usa.com    acme-admin3
```

**17.** Reboot the physical admin nodes, and watch the reboot on the console.

Wait for the physical admin nodes to come back up.

**18.** Log into the admin nodes as the root user.

**19.** Enter a `ping` command to another server at your site to make sure that the network is functioning.

Wait for the `ping` to return.

**20.** Log out of the console window, and use the `ssh` command to log into the admin node as the root user.

**21.** (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.

# (Conditional) Configuring the storage unit

Complete this procedure if you want to configure a system admin controller high availability (SAC HA) admin node.

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 output in the preceding steps indicates 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, preserve the data now by moving 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.

# (Conditional) Enabling an input-output memory management unit (IOMMU)

Complete this procedure if the following are both true:

- You want to configure a system admin controller high availability (SAC HA) admin node.
- The physical admin nodes are Intel platform admin nodes such as HPE Proliant DL360 servers.

**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
   ```

# Verifying the configuration

**Procedure**

1. Log into each physical admin node as the root user.

Complete the steps in this procedure on all physical admin nodes.

2. Enter the following command to verify the time zone:

   ```
   # date
   ```

3. Enter the following command to verify the hostname and the IP address:

   ```
   # cat /etc/hosts
   ```

4. Enter the following command to verify the time:

   ```
   # chronyc sources -v
   ```

5. Enter the following command to verify the network configuration:

   ```
   # ip addr
   ```

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.X (Ootpa)
          CPE OS Name: cpe:/o:redhat:enterprise_linux:8.X:GA
               Kernel: Linux 4.18.0-80.el8.x86_64
         Architecture: x86-64
   ```

7. (Conditional) Verify that IOMMU is enabled.

   Complete this step if you completed the following procedure:

   **(Conditional) Enabling an input-output memory management unit (IOMMU)**

   Enter the following command:

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

   The output is as follows on a correctly configured system:

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

8. (Conditional) Verify that the physical admin nodes can communicate with each other.

   Complete this step on physical admin nodes configured for high availability (HA).

   For admin nodes configured as system admin controller high availability (SAC HA) nodes, enter a `ping` command from physical admin node 1 to physical admin node 2. Also enter a `ping` command from physical admin node 2 to physical admin node 1.

   For admin nodes configured as quorum high availability HA nodes, enter commands as follows:

   - Enter a `ping` command from physical admin node 1 to the following:

    - Physical admin node 2
    - Physical admin node 3

- Enter a `ping` command from physical admin node 2 to the following:
    - Physical admin node 1
    - Physical admin node 3

- Enter a `ping` command from physical admin node 3 to the following:
    - Physical admin node 1
    - Physical admin node 2

# Slots

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 two.

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.

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.<br><br>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. |

# (Optional) Configuring a quorum high availability (quorum HA) admin node

A quorum HA admin node requires three physical admin nodes that use the x86_64 architecture. It uses the Gluster file system in sharding mode to host a virtual machine image. It uses Pacemaker to start and position the virtual machine as needed.

When the cluster is running, the admin node resides in a virtual machine upon one of three physical admin nodes. When a failover occurs, the virtual machine passes from the active node to one of the passive nodes.

The following procedures explain how to configure a quorum HA admin 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:

   **HPE Performance Cluster Manager operating system releases supported**
3. Enter the following command to create a directory for the installation `.iso` files:

   ```
   # mkdir /root/sw
   ```
4. Use your site practices to copy the installation `.iso` files from the server that hosts the installation files to the physical admin node.

   Example 1. Log into the host server and use the `scp` command as follows:

   ```
   scp host_system:/path/cm-admin-install-1.6-os-x86_64.iso /root/sw
   ```

   For *host_system*, specify the fully qualified domain name (FQDN) of the server that hosts the cluster manager `.iso` files.

   Example 2. Use the `cd` and `wget` commands:

   ```
   # cd /root/sw
   # wget http://host_system/path
   ```

   The variables are as follows:

   | Parameter | Specification |
   |---|---|
   | *host_system* | The FQDN of the server that hosts the cluster manager `.iso` files. |
   | *path* | The path to the `.iso` file on the server that resides on your corporate network. |
5. Open the following file in a text editor:

   `/opt/clmgr/etc/hadb.conf`

   The fields in this file are as follows:

| Field name | Information to provide |
|---|---|
| `phys1_hostname` | The hostname that you specified when you installed the operating system on this node. |
| `phys1_house_ip` | The IP address that you specified when you installed the operating system on this node. |
| `phys2_hostname` | The hostname that you specified when you installed the operating system on this node. |
| `phys2_house_ip` | The IP address that you specified when you installed the operating system on this node. |
| `phys3_hostname` | The hostname that you specified when you installed the operating system on this node. |
| `phys3_house_ip` | The IP address that you specified when you installed the operating system on this node. |
| `phys1_head_ip` | By default, this is set to `172.23.255.150`. |
| `phys2_head_ip` | By default, this is set to `172.23.255.151`. |
| `phys3_head_ip` | By default, this is set to `172.23.255.152`. |
| `head_netmask` | By default, this is set to `255.255.0.0`. |
| `predictable_network_support` | By default, this is set to `yes`. |
| `phys1_mgmt_nic1_ifname` | Log into physical node 1, and enter the following command to retrieve this information:<br><br>`ip addr show` |
| `phys1_mgmt_nic2_ifname` | Log into physical node 1, and enter the following command to retrieve this information:<br><br>`ip addr show` |
| `phys2_mgmt_nic1_ifname` | Log into physical node 2, and enter the following command to retrieve this information:<br><br>`ip addr show` |

*Table Continued*

| Field name | Information to provide |
| --- | --- |
| phys2_mgmt_nic2_ifname | Log into physical node 2, and enter the following command to retrieve this information:<br><br>`ip addr show` |
| phys3_mgmt_nic1_ifname | Log into physical node 3, and enter the following command to retrieve this information:<br><br>`ip addr show` |
| phys3_mgmt_nic2_ifname | Log into physical node 3, and enter the following command to retrieve this information:<br><br>`ip addr show` |
| phys1_bmc_ip | Specify the node controller IP address. Contact your network administrator regarding administrative and security requirements. |
| phys1_bmc_user | Specify the node controller username. Contact your network administrator regarding administrative and security requirements. |
| phys1_bmc_password | Specify the node controller password. Contact your network administrator regarding administrative and security requirements. |
| phys1_bmc_hostname | Specify the node controller hostname. Contact your network administrator regarding administrative and security requirements. |
| phys2_bmc_ip | Specify the node controller IP address. Contact your network administrator regarding administrative and security requirements. |
| phys2_bmc_user | Specify the node controller username. Contact your network administrator regarding administrative and security requirements. |
| phys2_bmc_password | Specify the node controller password. Contact your network administrator regarding administrative and security requirements. |
| phys2_bmc_hostname | Specify the node controller hostname. Contact your network administrator regarding administrative and security requirements. |
| phys3_bmc_ip | Specify the node controller IP address. Contact your network administrator regarding administrative and security requirements. |
| phys3_bmc_user | Specify the node controller username. Contact your network administrator regarding administrative and security requirements. |

*Table Continued*

| Field name | Information to provide |
|---|---|
| phys3_bmc_password | Specify the node controller password. Contact your network administrator regarding administrative and security requirements. |
| phys3_bmc_hostname | Specify the node controller hostname. Contact your network administrator regarding administrative and security requirements. |
| skip_firewall=no | By default, this is set to no. |
| phys1_head_hostname | Specify a name for the head network.<br>For example: acme-admin1-head |
| phys2_head_hostname | Specify a name for the head network.<br>For example: acme-admin2-head |
| phys3_head_hostname | Specify a name for the head network.<br>For example: acme-admin3-head |
| admin_iso_path= | Specify the path to the installation .iso file. The format is as follows:<br>admin_iso_path=/root/sw/cm-admin-install-*version*-*op_sys*-x86_64.iso<br>For example:<br>admin_iso_path=/root/sw/cm-admin-install-1.6-sles15sp3-x86_64.iso |

6. Enter the following command, and respond to the prompts:

   # **/opt/clmgr/lib/q-ha/setup**

   After the setup script in this step runs, the high availability admin node is created.

7. Use the ssh command to log into the third physical admin node, and enter the following command:

   # **virsh console adminvm**

8. Use one of the following procedures to install an operating system on the virtual machine:

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

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_monCluster Summary:
  * Stack: corosync
  * Current DC: nano-3 (version 2.0.5+20201202.ba59be712-2.30-2.0.5+20201202.ba59be712) - partition with quorum
  * Last updated: Mon Oct 18 08:20:10 2021
```

```
  * Last change:  Mon Sep 27 15:02:54 2021 by root via crm_resource on nano-2
  * 3 nodes configured
  * 4 resource instances configured

Node List:
  * Online: [ nano-1 nano-2 nano-3 ]

Active Resources:
  * p_ipmi_fencing_1    (stonith:external/ipmi):         Started nano-3
  * p_ipmi_fencing_2    (stonith:external/ipmi):         Started nano-1
  * p_ipmi_fencing_3    (stonith:external/ipmi):         Started nano-2
  * adminvm     (ocf::heartbeat:VirtualDomain):  Started nano-3
```

# (Optional) Configuring a system admin controller high availability (SAC HA) admin node

A SAC HA admin node requires two physical admin nodes that use the x86_64 architecture. When the cluster is running, the admin node resides in a virtual machine upon one of two physical admin nodes. When a failover occurs, the virtual machine passes from the active node to the passive node.

When you create a SAC 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.

---

**NOTE:** The cluster manager also supports the quorum HA admin node solution. It features three admin nodes, and it uses the Gluster file system in sharding mode to host a virtual machine image. It uses Pacemaker to start and position the virtual machine as needed.

The quorum HA solution eliminates the need for the HPE Modular Smart Array 2050 shared storage system that is required by the system admin controller (SAC) HA solution.

Certain clusters require a quorum HA admin node. For more information, see the following:

**(Optional) Configuring a quorum high availability (quorum HA) admin node**

---

The following procedures explain how to configure a SAC HA admin node:

**Procedure**

1. **Creating and installing the HA software repositories on the physical admin nodes**
2. **Preparing to run the HA admin node configuration script**
3. **Running the highly available (HA) admin node configuration script**
4. **Starting the HA virtual manager and installing the cluster manager on the virtual machine**

## 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:

   **HPE Performance Cluster Manager operating system releases supported**
3. Use the `ssh` command to log into the other admin node.

   When prompted, provide the root user login and password credentials.
4. 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/
```

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

   You need this information for the `admin_iso_path=` variable in the `sac-ha-initial-setup.conf` file.

   Enter the following commands:

   ```
   # mkdir /root/sw
   # ssh phys_admin2
   # mkdir /root/sw
   # scp host_system:/path/cm-admin-install-1.6-os-x86_64.iso /root/sw
   # rsync -avz /root/sw/ phys_admin:/root/sw/
   # exit
   ```

   The variables are as follows:

| Variable | Specification |
|---|---|
| *host_system* | The name of the node that currently hosts the `.iso` file. |
| *path* | The path to the `.iso` file on the host node. |
| *os* | 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-1.6-rhel84-x86_64.iso /root/sw/
   ```

# 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. On physical node 1, enter the `ip addr` command.

   The command displays NIC and MAC addresses. For example:

   ```
   linux:~ # ip addr
   1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
       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: eno1: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether ec:eb:b8:89:f2:90 brd ff:ff:ff:ff:ff:ff
3: eno2: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether ec:eb:b8:89:f2:91 brd ff:ff:ff:ff:ff:ff
4: eno3: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether ec:eb:b8:89:f2:92 brd ff:ff:ff:ff:ff:ff
5: eno5: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether 48:df:37:66:c1:30 brd ff:ff:ff:ff:ff:ff
6: eno4: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether ec:eb:b8:89:f2:93 brd ff:ff:ff:ff:ff:ff
7: eno6: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN group default qlen 1000
    link/ether 48:df:37:66:c1:38 brd ff:ff:ff:ff:ff:ff
linux:~ #
```

The interface names and the MAC addresses are highlighted in **bold** in the preceding output. In a subsequent step, you need this bolded information to specify the following:

- The physical MAC addresses for physical node 1
- Whether this node uses predictable network names

4. On physical node 2, enter the `ip addr` command.

   Again, you need the interface names and the MAC addresses in the command output to specify the following:

   - The physical MAC addresses for physical node 2
   - Whether this node uses predictable network names

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

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

   The installer automatically 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.

6. 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 4: 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 4: Configuration file inputs** contains information about the fields that you must edit. Do not edit the other fields.

   ---

**Table 4: Configuration file inputs**

| Configuration file line | Specification |
|---|---|
| Physical node 1 - MAC addresses: | |
| `phys1_eth0=""` | The MAC address shown in the output you retrieved from the command in Step **3**. |
| `phys1_eth1=""` | The MAC address shown in the output you retrieved from the command in Step **3**. |
| `phys1_eth2=""` | The MAC address shown in the output you retrieved from the command in Step **3**. |
| `phys1_eth3=""` | The MAC address shown in the output you retrieved from the command in Step **3**. |
| Physical node 2 - MAC addresses: | |
| `phys2_eth0=""` | The MAC address shown in the output you retrieved from the command in Step **4**. |
| `phys2_eth1=""` | The MAC address shown in the output you retrieved from the command in Step **4**. |
| `phys2_eth2=""` | The MAC address shown in the output you retrieved from the command in Step **4**. |
| `phys2_eth3=""` | The MAC address shown in the output you retrieved from the command in Step **4**. |
| Predictable network names for the physical nodes: | |

*Table Continued*

| Configuration file line | Specification |
|---|---|
| `predictable_network_support="yes"` | If physical node 1 and physical node 2 use predictable network names, do not edit this line.<br><br>For guidance, refer to the output from the following steps earlier in this procedure:<br><br>• Step **3**<br>• Step **4**<br><br>The nodes use predictable network names if the output shows interfaces with `enoX`. In this case, do not edit this line. That is, retain `predictable_network_support="yes"` in the file.<br><br>The nodes do not use predictable network names if the output shows interfaces with `ethX`. In this case, edit this line to appear as follows:<br><br>`predictable_network_support="no"` |
| `phys1_initial_house="eth0"` | If physical node 1 uses predictable network names, edit this line, and set this value to `eno1`.<br><br>If physical node 1 does not use predictable network names, do not edit this line.<br><br>**NOTE:** The value `eno1` includes a lowercase `o`, not a zero (`0`). |
| `phys2_initial_house="eth0"` | If physical node 2 uses predictable network names, edit this line, and set this value to `eno1`.<br><br>If physical node 2 does not use predictable network names, do not edit this line.<br><br>**NOTE:** The value `eno1` includes a lowercase `o`, not a zero (`0`). |
| 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. |

*Table Continued*

| Configuration file line | Specification |
|---|---|
| `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. |
| `bonding_method="active-backup"`<br>Or<br>`bonding_method="802.3ad"` | Bonding method to use for the `bond0` interface. Valid choices are `active-backup` or `802.3ad`. |
| Physical node 1 - Node controller hostname and node controller IP address: | |
| `phys1_bmc_hostname=""` | Obtain the node controller hostname for physical node 1 from your network administrator. |
| `phys1_bmc_ipaddr=""` | Obtain the node controller IP address for physical node 1 from your network administrator. |
| Physical node 2 - Node controller hostname and node controller IP address: | |
| `phys2_bmc_hostname=""` | Obtain the node controller hostname for physical node 2 from your network administrator. |
| `phys2_bmc_ipaddr=""` | Obtain the node controller IP address for physical node 2 from your network administrator. |
| Additional information: | |
| `wwn=""` | Specify the information you retrieved in the following procedure:<br>**(Conditional) Configuring the storage unit**<br>The value you need consists of a number string. For example:<br>`wwn="60080e5000233c340000039f4d90ab57"` |
| `volume_group_ID=""` | The number of the LUN on the storage device. Typically, this value is `0`.<br>For information about how to derive this number, see your storage unit documentation. |

*Table Continued*

| Configuration file line | Specification |
|---|---|
| `admin_iso_path=""` | The full path to the installation `.iso` file.<br><br>Specify the path that you configured in the following procedure:<br><br>**Creating and installing the HA software repositories on the physical admin nodes** |
| `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.<br><br>The following command retrieves CPU information for the `cpus=` field in the configuration file:<br><br>`# less /proc/cpuinfo`<br><br>In the output, look for the values for the following fields:<br><br>• `processor`<br>• `cpu cores`<br><br>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.<br><br>For example, set `cpus="6"` or `cpus="8"`. |
| `memory=""` | The amount of memory allocated to the virtual machine admin node that manages the cluster.<br><br>The following command retrieves information for the `memory=` field in the configuration file:<br><br>`# free -h`<br>`total . . .`<br>`Mem: 62G . . .`<br>`Swap: 2.0G . . .`<br><br>In the output, observe the values under the `total` column for the `Mem` field and the `Swap` field.<br><br>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.<br><br>For example, if you specified `cpus="6"`, specify `memory="24GB"`. If you specified `cpus="8"`, specify `memory="32GB"`. |

*Table Continued*

| Configuration file line | Specification |
|---|---|
| `rootsize=""` | Specify the amount of shared disk space on the storage unit that you want to allocate to the admin node virtual machine.<br><br>The `fdisk` command retrieves information for the `rootsize=` field in the configuration file. This command has the following format:<br><br>`fdisk -l` *`disk`*<br><br>For *disk*, specify the disk identifier for the shared disk.<br><br>For example:<br><br>`# fdisk -l /dev/sdb`<br>`Disk /dev/sdb: 4000.0 GB, . . .`<br>`.`<br>`.`<br>`.`<br><br>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.<br><br>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.

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

# Running the highly available (HA) admin node configuration script

The configuration script configures the two physical 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.

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

**Troubleshooting an HA admin node configuration**

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

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

   To obtain help information for the `sac-ha-initial-setup.conf` script, enter the following command:

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

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

   ```
   # reboot
   ```

   Wait for the boot to complete.

5. Log back into both of physical admin nodes as the root user.
6. 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
   /dev/mapper/sac-ha-vol3_mpath-part2  3.7T  146G  3.5T   4%   /images
   ```

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

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

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

8. (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
   Cluster Summary:
     * Stack: corosync
     * Current DC: hikari2 (version 2.0.5+20201202.ba59be712-2.30-2.0.5+20201202.ba59be712) - partition with quorum
     * Last updated: Wed Oct 20 10:15:39 2021
     * Last change:  Fri Sep 24 12:25:17 2021 by hacluster via crmd on hikari
     * 2 nodes configured
     * 9 resource instances configured

   Node List:
     * Online: [ hikari hikari2 ]

   Active Resources:
     * p_ipmi_fencing_1    (stonith:external/ipmi):         Started hikari2
     * p_ipmi_fencing_2    (stonith:external/ipmi):         Started hikari
     * sbd_stonith (stonith:external/sbd):  Started hikari2
     * Clone Set: dlm-o2cb-fs-images-clone [dlm-o2cb-fs-images-group]:
       * Started: [ hikari hikari2 ]
     * virt        (ocf::heartbeat:VirtualDomain):  Started hikari2
     * mailTo      (ocf::heartbeat:MailTo):         Started hikari2
   ```

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

   For example:

   ```
   # ls -lh /images/vms/
   total 135G
   -rw------- 1 qemu qemu 1.0T Sep 24 12:20 sac.img
   ```

# 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 in one of the following ways:

   - Log into the physical console. Make sure that the cluster is booted at run level 5 (`init 5`).

     Or

   - 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.

     Or

   - Log in through a VNC session. 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:

   a. Click the **Connect to remote host** box.

   b. In the **Hostname** field, enter the hostname of physical node 2.

   c. Click the **Autoconnect** box.

   d. 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, enter **`console=ttyS0,115200n8`**, and press Enter.

    This step enables you 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.

# Configuring the cluster software on the admin node

**Procedure**

1. **Preparing to configure the cluster software on the admin node**
2. Configuring the cluster software on the admin node. Use one of the following methods:
   - **Using the cluster definition file to specify the cluster configuration**

     Or
   - **Using the menu-driven cluster configuration tool to specify the cluster configuration**
3. **Completing the admin node software installation**
4. **(Conditional) Configuring an unsupported Ethernet switch into the cluster**
5. **(Optional) Configuring external domain name service (DNS) servers**

## Preparing to configure the cluster software on the admin node

**Procedure**

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

   You can configure the software from either physical media or from an ISO on your network.
2. From a graphics screen or through an `ssh` connection, log into the admin node as the root user., as follows:

   This step differs depending on whether your admin node is a single node or is a two-node SAC HA admin node, as follows:

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

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

See the following topic:

**Required ports**

# Required ports

| Service | Port(s) |
|---|---|
| External port for SSH | TCP 22 |
| External port for Kibana<br><br>If you want to use Kibana to access the centralized log files, TCP port 5601 is required. If necessary, you can change this port number after the cluster is configured. | TCP 5601 |
| External ports required for webpage and GUI. You can start the cluster manager web server on a different port. For more information, see the following:<br><br>**Starting the cluster manager web server on a non-default port** | TCP 80, 443, 1099, and 49150 |
| Internal ports required for monitoring: | UDP 48555 - 49587 |

For more information about port requirements, see the following:

- `/opt/clmgr/etc/cmuserver.conf`
- **HPE Performance Cluster Administration Guide**

# (Optional) Configuring the management network manually

The cluster manager installation process typically includes using a cluster definition file or running the `configure-cluster` command to configure the cluster management networks. As an alternative, you can use operating system commands to configure the management network manually. If you configure the management network manually, you can still use a cluster definition file or run the `configure-cluster` command to configure the rest of the cluster.

If you choose to configure the management network manually, observe the following requirements:

- Configure `bond0` with at least one IP address.
- To manage the node controllers, configure `bond0` with an additional IP address from the `head-bmc` management network. This is often noted as `bond0:bmc`.
- Combine IPV4 and IPV6 routes if needed.
- In the cluster definition file, you can set the management network interface with the following configuration attribute:

  `admin_mgmt_interfaces=existing`

  For example:

  .
  .

```
.
[attributes]
admin_house_interface=eno1
admin_mgmt_interfaces="existing"
admin_mgmt_bmc_interfaces="existing"
.
.
.
```

In the cluster definition file, you can also set the node controller network interfaces with the following configuration attribute:

```
admin_mgmt_bmc_interfaces=existing
```

- Make sure the cluster definition file describes the management network you configured. Failure to do so can produce unexpected results.

If you want to run the `configure-cluster` command after you configure the management network, navigate to the **Management Network Interfaces Selection** menu. That menu lets you specify network interfaces for `bond0` and lets you specify **Use existing Settings for Management**. Alternatively, make sure that the cluster definition file is complete, and supply the name of the cluster definition file as input to the `configure-cluster` command.

# Using the cluster definition file to specify the cluster configuration

**Prerequisites**

This method assumes that you have a cluster definition for the cluster.

**Procedure**

1. 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 physical 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-SP3-Full-x86_64-GM-Media1.iso
   # cm repo show
   ```

2. (Optional) Add updates and patches for the operating system software and for the cluster manager.

   Complete this step if updates are available and you want to update the software at this time.

   Cluster manager patch names and distribution update names can vary from these examples. The examples in this step add the repository as a custom repository and then select the patches. These commands assume the following:

   - The packages updates are at the following location:

     ```
     /opt/clmgr/repos/SLES15-SP3-Updates-x86_64
     ```

   - The cluster manager patches are in the following location:

```
/opt/clmgr/repos/patch11627-x86_64
```

Example 1. To add cluster manager `patch11627` on an x86_64 admin node, run the following commands:

```
# cm repo add /opt/clmgr/repos/patch11627-x86_64 --custom patch11627-x86_64
# cm repo select patch11627-x86_64
```

Example 2. To add SLES 15 SP3 update repos, run the following commands:

```
# cm repo add /opt/clmgr/repos/SLES15-SP3-Updates-x86_64 --custom SLES15-SP3-Updates-x86_64
# cm repo select SLES15-SP3-Updates-x86_64
```

3. 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.

# 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, complete the following steps:

   a. Use the space bar 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.

   b. Click **OK**.

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

   a. Use the space bar and arrow keys to select one or two NICs for the management network.

   b. 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 click **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 node controller traffic.

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

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

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

b. On the **Main Menu** screen, click **OK** to select the **Initial Setup Menu**.

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**.

   The next few steps describe how to create repositories for the following:

   - The operating system software for compute nodes and for infrastructure nodes
   - The cluster manager software
   - (Optional) Additional software for HPE Message Passing Interface (MPI), AMD ROCm, SLURM, or other products

   Locate your system disks before you proceed. The menu system prompts you to insert physical 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 **Would you like to create repos from media? ...** screen, select one of the following:

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

      Step **12**

Or

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

  Step **14**

12. 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 physical media, complete the following steps:

      a. Insert a DVD.

      b. Select **Mount inserted DVD**.

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

      d. On the **Would you like to create repos from media? ...** 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, complete the following steps:

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

      b. On the **Please enter the full path to the mount point or the ISO file ...** screen, 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.

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

      d. On the **Would you like to create repos from media? ...** 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.

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

    - Step **11**
    - Step **12**

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

14. 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.

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

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

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

18. Create additional networks for chassis management modules (CMMs) and chassis management controllers (CMCs).

By default, the cluster creates the `head` network and the `head-bmc` network. The `head` network and the `head-bmc` network are sufficient for clusters without CMMs or CMCs. Clusters with CMMs and CMCs need additional networks. Complete the following steps to create the CMM and CMC networks:

**a.** On the **Cluster Network Settings** screen, select one of the following:

- If the cluster is an HPE Cray EX cluster, select **L Create Default Cray EX Networks**.
- If the cluster is an HPE Apollo 9000, select **B Create Default Apollo 9000 Networks**.

**b.** On the **The default networks needed for *model_name* systems will be created** popup window, click **OK**.

**c.** On the **The *model_name* networks have been created** popup window, click **OK**.

**19.** Create one or more data networks.

Complete this step if the cluster needs additional data networks.

**NOTE:** The cluster manager requires each defined subnet address to be unique. That is, the head network and the BMC network cannot be the same.

The following substeps show how to create a data network and an InfiniBand network.

To configure a data network, complete the following steps:

**a.** On the **Cluster Network Settings** screen, select **A Add Subnet**, and select **OK**.

**b.** On the **Select network type** screen, press the space bar to move the asterisk (`*`) to the second line. This action selects the lower line, which now appears as follows:

```
(*) 4 Data Network
```

**c.** Select **OK**.

**d.** On the **Insert network name, subnet and netmask** screen, enter 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:<br>`data10g`. |
| **subnet** | The network IP address (start of the range) for the nodes on the data network. |
| **netmask** | Subnet mask for the nodes on the data network. |
| **gateway (optional)** | An IP address within the subnet that can be used as a default gateway. (Optional) |

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

To configure an InfiniBand network, for any kind of cluster, complete the following steps:

a. On the **Cluster Network Settings** screen, select **A Add Subnet**, and select **OK**.

b. On the **Select network type** screen, press the space bar to move the asterisk (*) to the second line. This action selects the lower line, which now appears as follows:

```
(*) 5 IB Network
```

c. Select **OK**.

d. On the **Insert network name, subnet and netmask** screen, enter information to define the InfiniBand 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 InfiniBand network. For example:<br>`ib0`. |
| **subnet** | The network IP address (start of the range) for the nodes on the data network. |
| **netmask** | Subnet mask for the nodes on the data network. |
| **gateway (optional)** | An IP address within the subnet that can be used as a default gateway. (Optional) |

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

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**.

      Or

    - To change the network settings, complete the following steps:

      a. Highlight the setting you want to change, and select **OK**.

      b. Enter a new subnet IP address, netmask, gateway, or VLAN, and select **OK**.

      c. 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 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**.

---

**NOTE:** For information about how to install the fabric management node (FMN), see the *HPE Cray Slingshot Operations Guide*.

---

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 `cm.clusterdomain.com`.
25. On the **Domain name configured** screen, click **OK**.
26. On the **Please adjust the domain_search_path as needed ...** screen, click **OK**.

    The default search paths use *head* and *head*`-BMC` networks. You can adjust this as needed after the cluster is configured. For information, see the following:

    **Adjusting the domain name service (DNS) search order**
27. Select **P Domain Search Path** to verify the domain search path.
28. (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.
29. On the **Cluster Network Settings** screen, adjust the VLAN settings.

    - If the cluster is an HPE Cray EX cluster, it includes chassis controllers. Complete the following steps to update the management VLAN settings:

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

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

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

         - **Management VLAN Start:** (Default is `2001`.) Set this to `2000`.
         - **Management Control VLAN Start:** (Default is `3001`.) Set this to `3000`.

         While on this screen, review the other settings and adjust as needed. The other settings are as follows:

- **CMMs per Rack** (Default is `8`.)
- **Rack Start Number:** (Default is `1`.)

  The **Rack Start Number** should be set to a value that is equal to the lowest-numbered cabinet number in the cluster. For example, if the lowest cabinet number is `x1000`, set the **Rack Start Number** to `1000`.

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

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

- If the cluster is an HPE Apollo 9000 cluster, it includes chassis controllers. You can disable or adjust the auto-generated routed VLANs on the management network. You do not have to change any of the defaults, but Hewlett Packard Enterprise recommends that you review the settings. 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 is `2001`.)
  - **Management VLAN End:** (Default is `2999`.)
  - **Management Control VLAN Start:** (Default is `3001`.)
  - **Management Control VLAN End:** (Default is `3999`.)
  - **CMCs per Management VLAN:** (Default is `8`.)
  - **CMCs per Rack:** (Default is `4`.)
  - **Rack Start Number:** (Default is `1`.)
  - **Management Network Subnet Selection:** (Default is `rack-based`. Alternative is `next-available`.)

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

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

- 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 Management Network Routing Settings**, and click **OK**.

  **b.** On the **Management Network Routing Settings** screen, select **O OSPF VLAN Settings** , and click **OK**.

  **c.** On the **Change OSPF VLAN #, Network, Subnet Mask** screen, use the up and down arrows to highlight the field you want to specify, specify a value, and click **OK**.

- **OSPF VLAN # [2~4094]**
- **OSPF Base Network [X.X.X.X]**
- **OSPF Base Netmask [X.X.X.X]**
- When finished specifying new values, click **OK**.
- When finished, click **Back**.

**d.** On the **Management Network Routing Settings** screen, select **R RIP VLAN Settings**, and select **OK**.

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

- **RIP VLAN # [2~4094]**
- **RIP Base Network [X.X.X.X]**
- **RIP Base Netmask [X.X.X.X]**
- When finished specifying new values, click **OK**.
- When finished, click **Back**.

**f.** When finished, click **Back**.

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

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

**32.** 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.
```

**33.** In the **Please enter the Domain Search Path for this cluster** box, verify the information, adjust if needed, and click **OK**.

**34.** On the **Domain Search Path Configured** screen, click **OK**.

**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.

  Or

- 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.

Or

- If the display is correct, select **Yes**.

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

    The following shows a representation of this screen:

```
Create which images now?
     [*] default  Default flat compute node image (Required)
     [*] lead     Leader node (RLC) image (Required for ICE)
     [*] ice      ICE compute node image (Required for ICE)
     [ ] none     Skip image creation (only check this box)

                  <  OK  >              < Back >
```

    Create only the `default` image.

    **NOTE:** On clusters with 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.

    For clusters with SU leader nodes, use the arrow keys and the space bar to specify **default** and to clear the **lead** and **ice** fields. Specify only a **default** image.

    When the screen shows the images that you want to create, select **OK**. 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. If you do not want the installer to create any images, select **none**.

    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.

    **NOTE:** Cluster manager support for Ganglia and Nagios will be removed in a future release.

    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 cluster-wide 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 cluster-wide 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 node discovery commands (again) on that component. This action 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**.

# Completing the admin node software installation

The following procedure completes the admin node software installation.

**Procedure**

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

   The `cattr` 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. Correct any aspect of the installation that is incorrect.

   For example, the installation process typically creates a default compute node image for you. If that process fails, the `cattr` output does not display a default image name. In this case, create one manually. When the `cmcinventory` service runs during the installation, it searches for a default image with a name that adheres to a specific format. Name the image so that it includes the distribution name, `rhel` or `sles`, plus the operating system distribution release level. For example, `rhel8.4`, `sles15sp4`, or any other image name that includes only the distribution name and the release level. At a minimum, include the cluster manager repository and the distribution repository in the default image you create. You can include additional repositories. For more information about how to create an image, see the following:

   **HPE Performance Cluster Administration Guide**

3. (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:

| Veriable | Specification |
|---|---|
| *phys1_head_ip_value* | The value in the `sac-ha-initial-setup.conf` file for `phys1_head_ip` |
| *phys1_eth1_value* | 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 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:

| Variable | Specification |
|---|---|
| *phys2_head_ip_value* | The `sac-ha-initial-setup.conf` file for `phys2_head_ip` |
| *phys2_eth1_value* | 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 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
```

# (Conditional) Configuring an unsupported Ethernet switch into the cluster

Complete this procedure if the cluster includes any unsupported Ethernet switches.

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.

**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 each node receives its image 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. (Optional) Create entries for the unsupported switch in the cluster definition file.

   When switch entries appear in the cluster definition file, the admin node assigns an IP address to a DHCP request from the switch. These entries also enable the admin node to match a static IP address for the switch to the hostname for the switch.

   For an example entry, see the following:

   **Cluster definition file example - Entries for an unsupported switch**

   ---

   **NOTE:** After the cluster manager installation is complete, 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 run the node discovery commands, DHCP assigns supported switches an IP address. In this way, you can use the `ssh` command or Telnet to connect 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.

   ---

# (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:** To enable NAT, complete the procedure in this topic at this time. You cannot complete the procedure in this topic after you run the node discovery commands. If you attempt to configure this feature after you run the node discovery commands, 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 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**.

# Verifying and splitting the cluster definition file

A **cluster definition file** contains the following:

- A list of cluster components
- Component-specific characteristics that need to be specified

Complete the following procedure to verify whether you have a cluster definition file and whether that 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.
   - If necessary, you can obtain the cluster definition file used in the manufacturing process from your technical support representative.

2. Split the cluster definition file into additional files.

| Cluster type | Content of the split files |
| --- | --- |
| HPE Cray EX with scalable unit (SU) leader nodes.<br><br>HPE Apollo 9000 with scalable unit (SU) leader nodes. | For these clusters, the original cluster definition file contains the information necessary to configure the management switches and the SU leader nodes.<br><br>Split the original file into two files, one for the management switches and one for the SU leader nodes.<br><br>If you have components such as power distribution units (PDUs) or additional compute nodes deployed as service nodes, create an additional file for these additional components.<br><br>The chassis controllers in these clusters facilitate automatic compute node configuration. Later in the installation process, you run a command to configure the compute nodes. |
| HPE Apollo cluster with SU leader nodes that is not an HPE Apollo 9000. | For these clusters, the original cluster definition file contains the information necessary to configure the management switches, the SU leader nodes, and the compute nodes.<br><br>Split the cluster definition file into three files, one for the management switches, one for the SU leader nodes, and one for the compute nodes.<br><br>If you have components such as power distribution units (PDUs) or additional compute nodes deployed as service nodes, create an additional file for these additional components. |
| Any cluster that requires a granular configuration approach. | Using more than one cluster definition file lets you take a step-by-step approach to the order in which components are configured into the cluster. Always create a cluster definition file for the management switches and use that file to configure the management switches into the cluster ahead of any other components.<br><br>Some clusters with leader nodes have compute nodes that serve as service nodes or login nodes and are attached directly to the admin node. Create an additional cluster definition file for those compute nodes, too. |

The following are example cluster definition file names and content:

| Example file name | Content |
|---|---|
| `mgmtsw.config` | Management switches only |
| `suleader.config` | SU leader nodes |
| `compute.config`<br>`pdu.config` | Compute nodes or components that the node discovery commands and configuration process do not configure automatically. |

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 one or two files.

**b.** Name the files according to the components they describe. For example, name one file for switches, one file for compute nodes, and (if present) one file for scalable unit (SU) leader nodes.

**c.** Open the file(s) you just created, and search for the `[discover]` section. Use an editor such as `vim`.

**d.** 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.

**e.** Review these files carefully before proceeding.

The following examples show the contents of example files. 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, include the following types of lines:

```
[discover]
internal_name=mgmtsw0, type=spine, ...
internal_name=mgmtsw1, type=leaf, ...
```

Example 2:

If a switch is not defined in the cluster definition file, enter information into the cluster definition file manually. To find the switch MAC address, either open a console to the switch or visually inspect the outside of the switch to find its label. If the switch does not support DHCP, configure a static IP address for the switch that matches the `mgmt_net_ip=` attribute in the configuration file. For example:

```
[discover]
# Aruba VSX Dual Control Plane Spine Management Switches
internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d4:43:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.252, hostname1=sw-spine01,
mgmtsw_partner=sw-spine02
internal_name=mgmtsw1, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d3:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.253, hostname1=sw-spine02,
mgmtsw_partner=sw-spine01
# Aruba VSX Dual Control Plane Leaf Management Switches
internal_name=mgmtsw2, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:ab:44:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.100, hostname1=sw-leaf01,
mgmtsw_partner=sw-leaf02
internal_name=mgmtsw3, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:cd:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.101, hostname1=sw-leaf02,
mgmtsw_partner=sw-leaf01
```

Example 3. To create a cluster definition file for compute nodes only, include the following types of lines:

```
[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, ...
```

Example 3. To create a cluster definition file for SU leader nodes, include the following types of lines:

```
[discover]
internal_name=service0, hostname1=leader1, mgmt_net_name=head, ...
internal_name=service1, hostname1=leader2, mgmt_net_name=head, ...
internal_name=service2, hostname1=leader3, mgmt_net_name=head, ...
```

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**

# 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 node discovery commands, specify the cluster definition file that includes the nodes and components that you want to configure.

The following table shows the types of cluster components in the cluster definition file:

| Component | Notes |
|---|---|
| Management switches | If you use management switches that HPE does not support, see the following before you include information about the switches in the cluster definition file:<br><br>**(Conditional) Configuring an unsupported Ethernet switch into the cluster**<br><br>**Configuring a new switch** |
| Power distribution units (PDUs) | The cluster manager does not configure PDUs into the cluster automatically on clusters with SU leader nodes. Define the PDUs in the cluster definition file.<br><br>PDUs are numbered starting with `0`. For example, `pdu0`, `pdu1`, `pdu2`, and so on. |

*Table Continued*

| Component | Notes |
| --- | --- |
| Scalable unit (SU) leader nodes | SU leader nodes are numbered starting with 1. For example, `leader1`, `leader2`, and so on. |
| Compute nodes | On HPE Cray EX clusters, the cluster manager uses a node template file. You can create a node template file with customized attributes, such as hostnames based on chassis location, for the compute nodes. Use the `cm node template` command to submit the template file to the cluster manager when it configures the compute nodes.<br><br>On HPE Apollo 9000 clusters, the cluster manager configures the compute nodes into the cluster automatically when the node discovery commands run. Do not include compute nodes in a cluster definition file. To assign site-specific hostnames to these nodes, assign the hostnames after the cluster is configured.<br><br>On HPE Apollo clusters other than the HPE Apollo 9000, include the compute nodes in a cluster definition file.<br><br>For information about required fields for compute nodes, enter the following command:<br><br>`# man cluster-configfile` |

In the cluster definition file, each component is defined with several configuration attributes. For example, these configuration attributes 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:

- 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`.

  Compute nodes with services installed upon them are numbered starting with 0. For example, `service0`, `service1`. These names are the default names for 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.
- 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 configuration attributes, 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
```

The node discovery commands include the `discover` command, the `cm node add` command, the `cm node discover add` command, and for some purposes, also the `configure-cluster` command. All these node discovery commands can accept a cluster definition file as input.

# 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 compute nodes and the leader nodes.

  Templates are useful when they pertain to multiple nodes, for example, many identical compute nodes. You can describe the nodes once, in the template section of the cluster definition file. The node template definitions can describe kernel names, image names, node controller 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 Cray EX cluster with scalable unit (SU) leader nodes

To configure an HPE Cray EX cluster, use one cluster definition file. For these systems, the cluster definition file defines the switches and the scalable unit (SU) leader nodes. In an HPE Cray EX cluster, a compute node that is part of an HPE Cray EX cabinet is configured into the cluster automatically through the `cmcinventory` service.

If the cluster has separate flat compute nodes, create a separate cluster definition file for those nodes. The example in this topic does not show this separate file.

Explanations for selected items are as follows:

- The `[templates]` section includes information that pertains to all of the SU leader nodes. The line that begins with the `name=su-leader` field 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.

The following is an example cluster definition file:

```
[templates]
name=su-leader, console_device=ttyS0, conserver_logging=yes, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc, rootfs=disk,
mgmt_net_bonding_master=bond0, dhcp_bootfile=grub2, disk_bootloader=no, mgmt_net_interfaces="ens2f0,ens2f1",
transport=udpcast, switch_mgmt_network=yes, tpm_boot=no, conserver_ondemand=no, mgmt_net_bonding_mode=802.3ad,
redundant_mgmt_network=yes, predictable_net_names=yes, baud_rate=115200, bmc_username=admin, bmc_password=admin,
image=su-leader, card_type=iLO

[nic_templates]
template=su-leader, network=head, bonding_master=bond0, bonding_mode=802.3ad, net_ifs="ens2f0,ens2f1"

[discover]
# admin node
internal_name=admin, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head, mgmt_net_macs="00:0f:53:3c:da:b0,00:0f:53:3c:da:b1",
mgmt_net_interfaces="ens2f0,ens2f1", mgmt_net_bonding_mode="802.3ad", admin_house_interface="eno1", hostname1=hog1,
rootfs=disk, redundant_mgmt_network=yes, predictable_net_names=yes, console_device=ttyS1,
architecture=x86_64

# Aruba Spine VSX Dual Control Plane Management Switches
internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d4:43:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.252, hostname1=sw-spine01,
mgmtsw_partner=sw-spine02

internal_name=mgmtsw1, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d3:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.253, hostname1=sw-spine02,
mgmtsw_partner=sw-spine01


# Aruba Spine VSX Dual Leaf Plane Management Switches
internal_name=mgmtsw2, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:ab:44:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.100, hostname1=sw-leaf01,
mgmtsw_partner=sw-leaf02

internal_name=mgmtsw3, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:cd:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.101, hostname1=sw-leaf02,
mgmtsw_partner=sw-leaf01

# Apollo 9K SU Leaders
internal_name=service1, hostname1=leader1, mgmt_bmc_net_macs="20:67:7c:e4:8a:8a",
mgmt_net_macs="3c:fd:fe:9c:ed:08,3c:fd:fe:9c:ed:09", 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="94:40:c9:47:20:02",
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


[dns]
cluster_domain=cm.clusterdomain.com
nameserver1=10.15.100.253
nameserver2=10.15.100.252

[attributes]
admin_house_interface=eno1
admin_mgmt_interfaces="ens2f0,ens2f1"
admin_mgmt_bmc_interfaces="ens2f0,ens2f1"
admin_udpcast_ttl=2
admin_udpcast_mcast_rdv_addr=239.255.255.1
admin_mgmt_bonding_mode=802.3ad
blademond_scan_interval=120
cmcs_per_mgmt_vlan=8
cmcs_per_rack=4
cmms_per_rack=8
conserver_logging=yes
conserver_ondemand=no
copy_admin_ssh_config=yes
dhcp_bootfile=grub2
discover_skip_switchconfig=no
domain_search_path=head.cm.clusterdomain.com,hostmgmt.cm.clusterdomain.com,
```

```
head-bmc.cm.clusterdomain.com,hostctrl.cm.clusterdomain.com,
cm.clusterdomain.com,clusterdomain.com
head_vlan=1
ipv6_local_site_ula=fdae:daac:11ba::/48
max_rack_irus=16
mcell_network=yes
mcell_vlan=3
mgmt_ctrl_vlan_end=3999
mgmt_ctrl_vlan_start=3000
mgmt_net_routing_protocol=ospf
mgmt_net_subnet_selection=rack-based
mgmt_vlan_end=2999
mgmt_vlan_start=2000
predictable_net_names=yes
rack_start_number=1000
rack_vlan_end=1100
rack_vlan_start=101
redundant_mgmt_network=yes
switch_mgmt_network=yes
udpcast_max_bitrate=900m
udpcast_max_wait=10
udpcast_mcast_rdv_addr=224.0.0.1
udpcast_min_receivers=1
udpcast_min_wait=10
udpcast_rexmit_hello_interval=0
monitoring_kafka_elk_alerta_enabled=no
monitoring_native_enabled=no

[networks]
name=public, subnet=129.111.3.0, netmask=255.255.255.0, gateway=129.111.3.1
name=head, type=mgmt, vlan=1, subnet=172.23.0.0, netmask=255.255.0.0, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0, netmask=255.255.0.0
name=hostctrl, type=mgmt-bmc, subnet=10.176.0.0, netmask=255.248.0.0, rack_netmask=255.255.252.0
name=hostmgmt, type=mgmt, subnet=10.168.0.0, netmask=255.248.0.0, rack_netmask=255.255.252.0
name=hsn, type=data, subnet=10.10.0.0, netmask=255.255.0.0

[images]
image_types="default"
```

## Cluster definition file example - HPE Apollo 9000 cluster with scalable unit (SU) leader nodes

To configure an HPE Apollo 9000 cluster, use one cluster definition file. For these systems, the cluster definition file defines the switches and the scalable unit (SU) leader nodes. In an HPE Apollo 9000 cluster, a compute node that is part of an HPE Apollo 9000 rack is configured into the cluster automatically through the `cmcinventory` service.

If the cluster has separate flat compute nodes, create a separate cluster definition file for those nodes. The example in this topic does not show this separate file.

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. The line that begins with the `name=su-leader` field 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
# /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-rhel8.4, 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, 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-rhel8.4
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-rhel8.4
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-rhel8.4

[dns]
cluster_domain=cm.clusterdomain.com
nameserver1=16.110.135.51
nameserver2=16.110.135.52

[attributes]
admin_house_interface=eno1
admin_mgmt_interfaces="ens2f0,ens2f1"
admin_mgmt_bmc_interfaces="ens2f0,ens2f1"
admin_udpcast_ttl=2
admin_udpcast_mcast_rdv_addr=239.255.255.1
admin_mgmt_bonding_mode=802.3ad
blademond_scan_interval=120
cmcs_per_mgmt_vlan=8
cmcs_per_rack=4
conserver_logging=yes
conserver_ondemand=no
copy_admin_ssh_config=yes
dhcp_bootfile=grub2
domain_search_path=head.cm.clusterdomain.com,hostmgmt.cm.clusterdomain.com,
head-bmc.cm.clusterdomain.com,hostctrl.cm.clusterdomain.com,cm.clusterdomain.com,
clusterdomain.com
head_vlan=1
max_rack_irus=16
mcell_network=no
mcell_vlan=3
mgmt_ctrl_vlan_end=3999
mgmt_ctrl_vlan_start=3001
mgmt_net_routing_protocol=ospf
mgmt_net_subnet_selection=rack-based
mgmt_vlan_end=2999
mgmt_vlan_start=2001
predictable_net_names=yes
rack_start_number=1
rack_vlan_end=1100
rack_vlan_start=101
redundant_mgmt_network=yes
switch_mgmt_network=yes
udpcast_max_bitrate=900m
udpcast_max_wait=10
udpcast_mcast_rdv_addr=224.0.0.1
udpcast_min_receivers=1
udpcast_min_wait=10
udpcast_rexmit_hello_interval=0
monitoring_kafka_elk_alerta_enabled=no
monitoring_native_enabled=no

[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, gateway=172.23.255.254
name=head-bmc, type=mgmt-bmc, vlan=1, subnet=172.24.0.0, netmask=255.255.0.0
name=hostctrl, type=mgmt-bmc, subnet=10.176.0.0, netmask=255.248.0.0, rack_netmask=255.255.252.0
name=hostmgmt, type=mgmt, subnet=10.168.0.0, netmask=255.248.0.0, rack_netmask=255.255.252.0
name=ib0, type=ib, subnet=10.148.0.0, netmask=255.255.0.0
name=ib1, type=ib, subnet=10.149.0.0, netmask=255.255.0.0

[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 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 with scalable unit (SU) leader nodes**

---

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, bmc_username=ADMIN, bmc_password=ADMIN,
force_disk="/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:0"

[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=ib0, net_ifs="ib0"
template=su-leader, network=ib1, 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 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=bt, 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=ib0, net_ifs="ib0"
template=su-compute, network=ib1, 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 - Virtual admin node on an HA admin cluster

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 cooling devices on an HPE Apollo 9000 cluster

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.
- 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 node discovery commands to put the cooling devices under cluster manager control.

## Cluster definition file example - Specifying a specific IP address

When you run the node discovery commands 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=ib0, net_ifs="ib0"
template=compute, network=ib1, 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:

**HPE Performance Cluster Administration Guide**

## 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 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=sles15sp3-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=ib0, net_ifs="ib0"
template=compute, network=ib1, 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 - HPE Apollo 20 nodes

If the cluster includes any HPE Apollo 20 compute nodes, set the `dhcp_bootfile=ipxe-direct` configuration attribute for these nodes in the cluster definition file. This attribute is required on HPE Apollo 20 compute nodes.

For example:

```
internal_name=service222, hostname1=apollo222, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="a4:bf:01:6a:08:73",
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, mgmt_net_macs="a4:bf:01:6a:08:72", disk_bootloader=no,
geolocation="apollo222", predictable_net_names=yes, mgmt_bmc_net_name=head-bmc, mgmt_net_name=head,
mgmt_net_bonding_master=bond0, transport=bt, redundant_mgmt_network=no, switch_mgmt_network=yes, dhcp_bootfile=ipxe-direct,
conserver_logging=yes, conserver_ondemand=no, tpm_boot=no, disk_bootloader=no, mgmtsw=mgmtsw0, console_device=ttyS0,
```

```
mgmt_net_bonding_mode=active-backup, rootfs=tmpfs, mgmt_net_interfaces="eno1", card_type=IPMI, bmc_username=admin123,
bmc_password=admin123, baud_rate=115200
```

## Cluster definition file example - HPE Apollo 80 nodes

This topic explains the following:

- How to check the cluster definition file for an HPE Apollo 80 node.
- How to complete the configuration for an HPE Apollo 80 node.

### Procedure

1. Create a cluster definition file that includes the chassis controller and the nodes.

   Key information in this file includes the `generic` keyword and the `mgmt_bmc_net_ip=` address. For example, assume that you create file `cmc.computes.config`with the following lines:

   ```
   # chassis controller info:
   [discover]
   temponame=service90, hostname1=a80cmc, generic, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="2c:d4:44:ce:ca:4c",
   mgmt_net_name=head, mgmt_net_macs="2c:d4:44:ce:ca:4b", mgmt_net_ip=172.23.0.15, mgmt_bmc_net_ip=172.24.0.15,
   switch_mgmt_network=yes, conserver_logging=no, conserver_ondemand=no

   # compute node info:
   [discover]
   internal_name=service2000, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="2c:d4:44:ce:c8:f8", mgmt_net_name=head,
   mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, mgmt_net_macs="2c:d4:44:ce:8a:2a",
   mgmt_net_interfaces="eno1", hostname1=baymax1, rootfs=disk, transport=udpcast, conserver_ondemand=no, tpm_boot=no,
   predictable_net_names=yes, dhcp_bootfile=grub2, conserver_logging=yes, disk_bootloader=no, switch_mgmt_network=yes,
   redundant_mgmt_network=no, console_device=ttyAMA0, architecture=aarch64, card_type=bmx, baud_rate=115200,
   bmc_username=hpcadmin, bmc_password=HPCADMIN, image=aarch64-rhel8.4, kernel=4.18.0-80.el8.aarch64,
   mgmt_bmc_net_ip=172.24.0.15, rack_nr=1, chassis=1, node_nr=0

   internal_name=service2001, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="2c:d4:44:ce:c8:f8", mgmt_net_name=head,
   mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, mgmt_net_macs="2C:D4:44:CE:8A:29",
   mgmt_net_interfaces="eno1", hostname1=baymax2, rootfs=disk, transport=udpcast, conserver_ondemand=no, tpm_boot=no,
   predictable_net_names=yes, dhcp_bootfile=grub2, conserver_logging=yes, disk_bootloader=no, switch_mgmt_network=yes,
   redundant_mgmt_network=no, console_device=ttyAMA0, architecture=aarch64, card_type=bmx, baud_rate=115200,
   bmc_username=hpcadmin, bmc_password=HPCADMIN, image=aarch64-rhel8.4, kernel=4.18.0-80.el8.aarch64,
   mgmt_bmc_net_ip=172.24.0.15, rack_nr=1, chassis=1, node_nr=1
   ```

2. Use the `cm node add` command in the following format to configure the components into the cluster:

   `cm node add -c config_file`

   For *config_file*, specify the name of the first cluster definition file you edited in this procedure. This file includes information about the chassis controllers and switches. For example, specify `cmc.computes.config`.

3. Use the `cm node provision` command to provision each node with an image.

4. Proceed to the following to complete the cluster configuration:

   **Backing up the cluster**

## Cluster definition file example - Entries for service nodes with NICs for a data network

If you used the menu-driven cluster configuration tool to create a data network, create a cluster definition file for the service nodes that host the data network.

Specify this cluster definition file to the `cm node add` command to configure these nodes into the cluster. Run the `cm node add` against this file before you configure the compute nodes into the cluster. Alternatively, you could include these nodes in the cluster definition file.

### Procedure

1. On the admin node, create a new cluster definition file, and add the node specifications for the two service nodes to the file.

The following shows a completed example cluster definition file for the two services nodes. The two data networks' attributes are shown in **bold**:

```
[discover]
hostname1=toki-1-srv0, internal_name=service0, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="0c:c4:7a:1b:45:93",
mgmt_net_name=head, mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup,
mgmt_net_macs="0c:c4:7a:14:04:6e", mgmt_net_interfaces="ens1f0", mgmt_net_interface_name="toki-1-srv0",
data1_net_name=ib0, data1_net_interfaces="ib0", data1_net_interface_name="toki-1-srv0-ib0", data2_net_name=ib1,
data2_net_interfaces="ib1", data2_net_interface_name="toki-1-srv0-ib1", rootfs=disk, transport=udpcast,
conserver_logging=yes, conserver_ondemand=no, dhcp_bootfile=grub2, disk_bootloader=no, mgmtsw=mgmtsw0,
predictable_net_names=yes, redundant_mgmt_network=no, switch_mgmt_network=yes, tpm_boot=no, console_device=ttyS1,
architecture=x86_64, card_type="IPMI"
```

**NOTE:** To configure service nodes for a high-speed network or a 10G network, create a cluster definition file similar to the one in this step.

2. Use the `cm node provision` command to provision each node with an image.

## Cluster definition file example - Attributes for a management switch

The cluster manager supports different types of redundancy protocol switches. Traditionally, the terminology for redundancy has been **stacking**, which means two or more physical switches act as a single logical switch. This is known as **single control plane**. When two physical switches each act as independent logical switches, this is known as **dual control plane**.

When you define a dual control plane spine or leaf type management switch, specify the `mgmtsw_partner=`*`hostname`* attribute in the cluster definition file to define the dual control plane partner switch.

The following example defines a dual control plane spine switch in the cluster definition file:

```
[discover]
internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d4:43:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.252, hostname1=sw-spine01,
mgmtsw_partner=sw-spine02
internal_name=mgmtsw1, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:d3:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-spine, mgmt_net_ip=172.23.255.253, hostname1=sw-spine02,
mgmtsw_partner=sw-spine01
```

**NOTE:** When you define the IP addresses of the dual control plane spine switches, do not specify the IP address of the head network gateway. The dual control plane switches use an Active Gateway protocol to emulate the head network gateway. This protocol virtualizes the IP address in case of partial switch failure.

The following command shows how to identify this IP address:

```
# cadmin --show-head-gateway
172.23.255.254
```

The following example defines a dual control-plane leaf switch in the cluster definition file:

```
[discover]
internal_name=mgmtsw2, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:ab:44:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.100, hostname1=sw-leaf01,
mgmtsw_partner=sw-leaf02
internal_name=mgmtsw3, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:cd:07:00", redundant_mgmt_network=yes,
net=head/head-bmc, ice=no, type=dual-leaf, mgmt_net_ip=172.23.255.101, hostname1=sw-leaf02,
mgmtsw_partner=sw-leaf01
```

The following example defines a single control plane spine switch in the cluster definition file:

```
internal_name=mgmtsw0, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:aa:32:12",
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=spine, mgmt_net_ip=172.23.255.254
```

The following example defines a single control plane leaf switch in the cluster definition file:

```
internal_name=mgmtsw1, mgmt_net_name=head, mgmt_net_macs="b8:d4:e7:ba:56:12",
redundant_mgmt_network=yes, net=head/head-bmc, ice=no, type=leaf, mgmt_net_ip=172.23.255.100
```

## Cluster definition file example - Entries for an unsupported switch

The following entries define an unsupported switch in the cluster definition file:

```
### Example of a config file with unsupported switches and a defined IP address
[discover]
internal_name=service50, hostname1=dell-sw1, mgmt_net_name=head, mgmt_net_macs="0a:cc:99:98:e5:af", generic,
mgmt_net_ip=172.23.255.240
internal_name=service51, hostname1=dell-sw2, mgmt_net_name=head, mgmt_net_macs="0a:cc:99:98:e7:aa", generic,
mgmt_net_ip=172.23.255.241
```

# (Optional) Creating a custom partitions configuration file

The procedure in this topic explains how to create a configuration file for custom partitions on one or more compute nodes.

**Prerequisites**

- **(Optional) Configuring custom partitions on the admin node**
- Respond to the installation dialog prompts in a way that facilitates custom partitions as described in the following topic:

  **Inserting the installation DVD and booting the admin node**

**Procedure**

1. Change to the following directory:

   `/opt/clmgr/image/scripts/pre-install`

2. 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.

3. Use the guidelines in the custom partition configuration file to describe the custom partitions you want to create.

4. Save and close the custom partition configuration file.

5. Open the cluster definition file for the compute nodes.

6. Add information about compute node custom partitions to the cluster definition file.

   Decide which compute nodes require custom partitions. Locate the node definition lines for those nodes in the compute node cluster definition file. For each line, add the following configuration attribute, which points to the custom partition configuration file:

   `custom_partitions=`*`file`*`.cfg`

   For example, assume that node `service1` uses the partition layout specified in the default custom partition file `custom_partitions.cfg`. You could have the following specification in the cluster definition file for compute nodes:

   ```
   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, conserver_logging=yes,
   conserver_ondemand=no, console_device=ttyS1,
   custom_partitions=custom_partitions.cfg
   ```

7. Save and close the cluster definition file.

# Configuring the management switches into the cluster

The procedure in this topic adds the management switches into the cluster.

**Procedure**

1. Log into the admin node as the root user.
2. Verify the cluster definition files you created.
3. Configure the management switches into the cluster.

   Use the `cm node add` command in the following format:

   `cm node add -c` *`management_switch_file`*

   For *management_switch_file*, specify the name of the file that includes information about the management switches.

   For example:

   ```
   # cm node add -c mgmtsw.config
   ```

4. (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
   ```

5. 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 are as follows:

   | Variable | Specification |
   |---|---|
   | *hostname* | The hostname of the management switch |
   | *new_password* | A strong, new password for the switch |

   For example:

   ```
   # switchconfig change_password --switches mgmtsw0 --new Hp3@dm!n2o20
   ```

   **NOTE:** Hewlett Packard Enterprise strongly recommends that you implement standard and secure practices to store all passwords at your site. Do not lose this information.

6. Enter the following command to save the changed configuration to the nonvolatile memory (NVM) on the switches:

   ```
   # switchconfig config -s all --save
   ```

# Configuring scalable unit (SU) leader nodes

Use the following procedures to configure the SU leader nodes and the compute nodes of a cluster with SU leader nodes.

**Procedure**

1. Familiarizing yourself with the files used to configure SU leader nodes. For information, see the following topic:

   **Files used when configuring scalable unit (SU) leader nodes**

2. **Creating a scalable unit (SU) leader node image**
3. Adding the SU leader nodes to the cluster. Use one of the following platform-specific procedures:

   - **Adding the scalable unit (SU) leader nodes to the cluster by using a cluster definition file**
   - **Adding the scalable unit (SU) leader nodes to the cluster without a cluster definition file**

4. **Configuring a static IP address for the node controllers (baseboard management controller (BMC) or iLO devices) of the scalable unit (SU) leader nodes**
5. **Configuring bonding on the scalable unit (SU) leader nodes**
6. **Determining the status of the scalable unit (SU) leader node list file**
7. **(Conditional) Creating a scalable unit (SU) leader node list file**
8. Configuring the SU leader node software. Use one of the following platform-specific procedures:

   - **Configuring the scalable unit (SU) leader node software on an HPE Cray EX cluster or an HPE Apollo 9000 cluster**
   - **Configuring the scalable unit (SU) leader node software on an HPE Apollo cluster with SU leader nodes that is not an HPE Apollo 9000 cluster**

9. (Conditional) Configuring the chassis controllers.

   Not all clusters have chassis controllers. Complete one of the following platform-specific procedures depending on your cluster type:

   - **Configuring the chassis controllers on an HPE Cray EX cluster**
   - **Configuring the chassis controllers on an HPE Apollo 9000 cluster**

10. **(Optional) Creating a chassis management controller (CMC) template file for an HPE Cray EX cluster**
11. Configuring the compute nodes. Use one of the following platform-specific procedures:

    - **Configuring the compute nodes into an HPE Cray EX cluster or an HPE Apollo 9000 cluster**
    - **Configuring the compute nodes into an HPE Apollo cluster with scalable unit (SU) leader nodes than is not an HPE Apollo 9000 cluster**

## 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.

- Two or more cluster definition files. You need one for management switches and one for SU leader nodes.

  For information about splitting the cluster definition file, see the following:

  **Verifying and splitting the cluster definition file**

  You can obtain a copy of the original version of this file from HPE if you need it in the future.

- `/opt/clmgr/etc/cmcinventory.conf`

  Depending on the cluster, you might want to change one or more settings in the `cmcinventory.conf` file. For example:

  - `db_write`. By default, the cluster manager updates the cluster manger internal database and adds new nodes.

    If you set `db_write = False`, you can simulate a dry-run of the installation and verify the changes that the cluster manager will make. The effect is as follows:

    - `cmcinventory` generates a `fastdiscover.conf` file.
    - `cmcinventory` does not update the database or any node settings in the database.
    - `cmcinventory` does not run any underlying commands.

  - `discover_add`. When you add new hardware, set `discover_add = False`. This setting directs `cmcinventory` to run the `fastdiscover` command to add new nodes found on the system chassis controllers.
  - `initial_power_on`. By default, the cluster manager powers on all hardware that it can detect. To suppress the power on, set `initial_power_on = False`.
  - `mac_update`.

    Hewlett Packard Enterprise recommends that you use the default of `mac_update = True`. With this setting, `cmcinventory` updates the MAC addresses in the database when a hardware swap has occurred. It leaves all other settings alone. The cluster database rules prevent duplicate MAC addresses from residing in the database. When the default setting is in effect, and `cmcinventory` detects a duplicate MAC address, it enters NULL for the MAC address of the old node location and then adds the MAC address back into the database in the new location.

    The `mac_update = False` setting prevents `cmcinventory` from updating MAC addresses in the cluster database. You, the user, take the responsibility to delete MAC addresses from the cluster database for removed or swapped nodes.

  - `soft_delete`.

    When you set `soft_delete = True`, the following occurs:

    - The cluster manager excludes information about removed hardware in command output.
    - The `cmcinventory` service sets the node to a non-exist (`18`) administrative state, which essentially removes the node from the cluster database without losing the specific settings for that node

By default, `soft_delete = False`, which leaves the administrative state as is. For example, if a node has being physically removed, its administrative state can be online. You can set the administrative state as you want.

For more information, see the following:

- The `cmcinventory` manpage.
- The `/opt/clmgr/etc/cmcinventory.conf` configuration file.

# Creating a scalable unit (SU) leader node image

The cluster manager requires an x86_64 architecture for SU leader nodes. If you need to reinstall the cluster software on a cluster with SU leader nodes, configure the same number of SU leader nodes that you had before.

**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 8.4:

   ```
   # cm image create -i su-rhel8.4 -l /opt/clmgr/image/rpmlists/generated/generated-rhel8.4.rpmlist
   ```

   For RHEL 7.9:

   ```
   # cm image create -i su-rhel7.9 -l /opt/clmgr/image/rpmlists/generated/generated-rhel7.9.rpmlist
   ```

   For SLES 15 SP3:

   ```
   # cm image create -i su-sles15sp3 -l /opt/clmgr/image/rpmlists/generated/generated-sles15sp3.rpmlist
   ```

   For SLES 12 SP5:

   ```
   # cm image create -i su-sles12sp5 -l /opt/clmgr/image/rpmlists/generated/generated-sles12sp5.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 8.4:

     ```
     # cm image dnf -i su-rhel8.4 install su-leader-collection
     ```

   - Enter the following command for RHEL 7.9:

     ```
     # cm image yum -i su-rhel7.9 install su-leader-collection
     ```

   - Enter the following command for SLES 15 SP3:

     ```
     # cm image zypper -i su-sles15sp3 install su-leader-collection
     ```

   - Enter the following command for SLES 12 SP5:

     ```
     # cm image zypper -i su-sles12sp5 install su-leader-collection
     ```

3. (Conditional) Add the `image=` configuration attribute to the SU leader information in the cluster definition file.

Complete this step if you have a cluster definition file.

Specify the SU leader node image you just created. For example, the following line specifies `image=su-leader` as the image for all SU leader nodes:

```
[templates]
name=su-leader, console_device=ttyS0, conserver_logging=yes, mgmt_net_name=head, mgmt_bmc_net_name=head-bmc,
rootfs=disk, mgmt_net_bonding_master=bond0, dhcp_bootfile=grub2, disk_bootloader=no,
mgmt_net_interfaces="ens2f0,ens2f1", transport=udpcast, switch_mgmt_network=yes, tpm_boot=no,
conserver_ondemand=no, mgmt_net_bonding_mode=802.3ad, redundant_mgmt_network=yes, predictable_net_names=yes,
baud_rate=115200, bmc_username=admin, bmc_password=admin, image=su-sles15sp3, card_type=iLO
```

For information about how to open a cluster definition file and how to edit the file, see the following:

**Verifying and splitting the cluster definition file**

# Adding the scalable unit (SU) leader nodes to the cluster by using a cluster definition file

Complete the procedure in this topic if you have a cluster with SU leader nodes and a cluster definition file.

**Procedure**

1. Log into the admin node as the root user.
2. Enter the following command:

   `cm node add [-n su_leaders] -c cluster_definition_file_for_new_nodes`

   For *cluster_definition_file_for_new_nodes*, specify the name of your cluster definition file.

   For example:

   ```
   # cm node add -n leader[1-30] -c suleader.config
   ```

3. Use the `cm node provision` command to provision the new SU leader nodes with an image and (optionally) to power cycle the new compute nodes.

   ```
   cm node provision -i su_image -n list_of_SU_leader_hostnames
   ```

   For the hostnames, can provide a range or a list of hostname.

   ```
   # cm node provision -i my.su.image -n leader[1-30]
   ```

   At this point, the SUs are configured into the cluster database, booted, and running.

# Adding the scalable unit (SU) leader nodes to the cluster without a cluster definition file

Use the `cm node discover` command to configure the SU leader nodes into the cluster when you do not have a cluster definition file and you do not know the MAC addresses of the SU leader nodes.

**Procedure**

1. Complete the procedure in the following topic:

   **Configuring compute nodes without a cluster definition file by using the `cm node discover` command**

2. Return here and continue with the SU leader node configuration process.

# Configuring a static IP address for the node controllers (baseboard management controller (BMC) or iLO devices) of the scalable unit (SU) leader nodes

In a cluster with SU Leaders, the admin node joins the SU leaders to participate in the storage. If the SU leaders are not brought up, then the shared paths on the admin node are not accessible. This situation can affect the DHCP service. If DHCP is unable to start or service the IP addresses for the node controllers, it is hard to reach the node controllers on the SU leader nodes to power up the SU leader nodes.

The procedure in this topic explains how to configure a static IP address on each node controller (the baseboard management controller (BMC) or iLO device) on each SU leader. With the static IP addresses, you can always reach the node controllers on the SU leader nodes to power them up.

A static IP address in these nodes facilitates the power-up process.

**Procedure**

1. Configure static IP addresses for the node controllers on the SU leader nodes.

   To accomplish this task, take the IP addresses that the cluster manager assigned to the SU leader nodes, and configure the node controller on each SU leader node to use that IP address statically and not rely on DHCP to supply it. The exact way to configure this static IP address varies depending on the type of node controller hardware.

2. Edit the node definitions in the cluster definition file to use the static IP addresses you assigned.

   Specify a `mgmt_bmc_net_ip=` configuration attribute for the node controller of each SU leader node.

   By editing the cluster definition file with this information, you ensure that the new addresses are used the next time there is a power failure or you decide to reinstall the cluster software.

3. (Optional) Configure the BIOS on the SU leader nodes and (possibly) on the admin node to **not** power on automatically if AC power is lost.

   Typically, the BIOS systems in infrastructure nodes, such as the admin node and the SU leader nodes, are configured to always power on after power is lost. In that case, the admin node and the SU leader nodes automatically power on when power is restored.

# Configuring bonding on the scalable unit (SU) leader nodes

**Procedure**

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

2. Use the `switchconfig_configure_node` command to configure the management switches to match the bonding mode on the SU leader nodes.

   The format is as follows:

   `switchconfig_configure_node --node` *`hostnames`*

   For *hostnames*, specify the SU leader node hostnames.

   For example:

   ```
   # switchconfig_configure_node --node leader[1-3]
   ```

# Determining the status of the scalable unit (SU) leader node list file

**Procedure**

1. Enter the following command to verify whether the node list file exists:

   ```
   # cat /opt/clmgr/etc/su-leader-nodes.lst
   ```

   If the file exists, proceed to Step **2** in this procedure.

   If the file does not exist, proceed to the following:

   **(Conditional) Creating a scalable unit (SU) leader node list file**

2. Assess the file.

   A usable file 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. Update the file as needed.

3. Back up the SU leader node list file.

   For example, enter the following command:

   ```
   # cp /opt/clmgr/etc/su-leader-nodes.lst /opt/clmgr/etc/su-leader-nodes.lst.BACKUP
   ```

# (Conditional) Creating a scalable unit (SU) leader node list file

If you do not have an SU leader node list file, you can obtain a copy of the original SU leader node list file from the HPE factory and update that file according to the instructions in this procedure.

The procedure in this topic explains how to create an SU leader node list file and write it to the correct location, which is as follows:

`/opt/clmgr/etc/su-leader-nodes.lst`

**NOTE:** Blank lines cannot appear in the `/opt/clmgr/etc/su-leader-nodes.lst` file.

**Procedure**

1. 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 node controller network for each SU leader node.

     This IP address enables SU leaders to communicate with the node controllers in the cluster. This is an alias on `bond0` on the node, and it is not the same as the IP address on the BMC/node controller for this node. It is unique.

   - An IP address for the pool of IP addresses used by compute nodes to reach SU leader nodes in a high availability manner. The SU leader cluster trivial database (CTDB) facilitates high availability and IP distribution.

     This IP address allows a Gluster resource to move to a different node. This IP address is on the head/management network, but it is different from the existing IP address. This IP address becomes an alias on `bond0` and is unique.

The IP addresses you choose cannot be the management network (`mgmt`) or management BMC network (`mgmt-bmc`) IP addresses of the SU leaders themselves. Instead, select unused IP addresses that are in the allowed ranges for the `mgmt` network and the `mgmt-bmc` network. After the configuration is complete, the SU leader nodes use these IP addresses to access the `mgmt` network and the `mgmt-bmc` network.

For example, consider the following specification in `/opt/clmgr/etc/su-leader-nodes.lst`:

```
leader1,172.24.255.241,172.23.255.241,/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:1
leader2,172.24.255.242,172.23.255.242,/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:1
leader3,172.24.255.243,172.23.255.243,/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:1
```

The preceding specification yields two additional aliases on `bond0`:

```
bond0: <BROADCAST,MULTICAST,MASTER,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP group
default qlen 1000 link/ether 48:df:37:bd:94:a0 brd ff:ff:ff:ff:ff:ff inet 172.23.100.11/16 brd 172.23.255.255
scope global bond0 valid_lft forever preferred_lft forever inet 172.24.255.241/16 brd 172.24.255.255 scope global
bond0:bmc valid_lft forever preferred_lft forever inet 172.23.255.243/16 brd 172.23.255.255 scope global
secondary bond0 valid_lft forever preferred_lft forever
inet6 fe80::4adf:37ff:febd:94a0/64 scope link valid_lft forever preferred_lft forever
```

If the default network range is used for the head and bmc network the cluster manager uses IP addresses `172.23.255.X` and `172.24.255.X`, respectively. Ensure that the IP addresses you choose are unique.

To verify that the IP address you chose are unique, enter the following commands:

```
# cm node show --ips -n '*' | grep head
# cm node show --ips -n '*' | grep head-bmc
```

2. Open `/opt/clmgr/etc/su-leader-nodes.lst` with a text editor and enter the two new, unique IP addresses in the second field and the third field.

   The line includes the following fields, separated with a comma:

   *`SU_leader_hostname`*`,`*`node_ctrlr_NIC_IP_address`*`,`*`mgmt_IP_addr`*`,`*`path_to_LUN`*

   Save and close the file after you complete your edits.

3. 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.

4. 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 **5**

- 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 **4.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 **5**

**5.** 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.

6. Back up the SU leader node list file.

   For example, enter the following command:

   ```
   # cp /opt/clmgr/etc/su-leader-nodes.lst /opt/clmgr/etc/su-leader-nodes.lst.BACKUP
   ```

# Configuring the scalable unit (SU) leader node software on an HPE Cray EX cluster or an HPE Apollo 9000 cluster

**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 *SU_lead_hostname*, specify the hostname of one of the SU leader nodes.

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
```

**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 `cm image activate` 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 in shared storage.

Enter the `cm image activate` command in the following format:

```
cm image activate -i image_name
```

For *image_name*, specify the image name.

For example:

```
# cm image activate -i rhel8.4
```

# Configuring the scalable unit (SU) leader node software on an HPE Apollo cluster with SU leader nodes that is not an HPE Apollo 9000 cluster

**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 *SU_lead_hostname*, specify the hostname of one of the SU leader nodes.

      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
```

**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 `cm image activate` 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 in shared storage.

Enter the `cm image activate` command in the following format:

`cm image activate -i image_name`

For *image_name*, specify the image name.

For example:

```
# cm image activate -i rhel8.4
```

# Configuring the chassis controllers on an HPE Cray EX cluster

**Procedure**

1. Gather the information about the cabinets and switches in the cluster.

   By default, chassis management modules (CMMs) are cabled to both of the switches in a cabinet. Also by default, chassis environment controllers (CECs) are cabled to the first switch in a cabinet.

   Enter the following command to display the switch hostnames:

   ```
   # cm group system show mgmt_switch
   ```

   The following additional information can be obtained through a visual inspection and the cluster cabling diagram. For the example in this step, assume the following pertains to an example cluster with default cabling:

   - The four cabinets are numbered `x1000` through `x1003`.

     The switches are a pair of Aruba 8360 switches in VSX mode. The switch hostnames are `sw-cdu01` and `sw-cdu02`.
   - The CECs map to the switches as follows:
     - Cabinet `x1000` CECs are cabled to `sw-cdu01` on ports 1/1/41,1/1/45.
     - Cabinet `x1001` CECs are cabled to `sw-cdu01` on ports 1/1/42,1/1/46.
     - Cabinet `x1002` CECs are cabled to `sw-cdu01` on ports 1/1/43,1/1/47.
     - Cabinet `x1003` CECs are cabled to `sw-cdu01` on ports 1/1/44,1/1/48.
   - The 32 CMMs map to the switches as follows:
     - Cabinet `x1000` CMMs are cabled to `sw-cdu01` and `sw-cdu02` on ports 1/1/1-8.
     - Cabinet `x1001` CMMs are cabled to `sw-cdu01` and `sw-cdu02` on ports 1/1/11-18.
     - Cabinet `x1002` CMMs are cabled to `sw-cdu01` and `sw-cdu02` on ports 1/1/21-28.
     - Cabinet `x1003` CMMs are cabled to `sw-cdu01` and `sw-cdu02` on ports 1/1/31-38.

2. Use the `switchconfig` command to configure the chassis environment controller (CEC) switch ports, and then restart the ports.

   Given the example configuration, enter the following commands to configure ports 41 through 48 and then restart:

   ```
   # switchconfig set -s sw-cdu01 --ports 1/1/41,1/1/45 --default-vlan 3000
   # switchconfig set -s sw-cdu01 --ports 1/1/42,1/1/46 --default-vlan 3001
   # switchconfig set -s sw-cdu01 --ports 1/1/43,1/1/47 --default-vlan 3002
   # switchconfig set -s sw-cdu01 --ports 1/1/44,1/1/48 --default-vlan 3003
   # switchconfig port -s sw-cdu01 --restart --ports 1/1/41-48
   ```

   For information about VLANs, see the following:

   **Subnetwork information**

3. Generate one `cmcdetectd` file for each cabinet.

   Given the example configuration, enter the following commands, which write information to the default file of `/etc/cmc-switch-info.txt`:

   ```
   # cmcdetectd --ex-generate --cabinet 1000 --mgmtsw sw-cdu01 --port-range 1/1/1-8
   # cmcdetectd --ex-generate --cabinet 1001 --mgmtsw sw-cdu01 --port-range 1/1/11-18
   # cmcdetectd --ex-generate --cabinet 1002 --mgmtsw sw-cdu01 --port-range 1/1/21-28
   # cmcdetectd --ex-generate --cabinet 1003 --mgmtsw sw-cdu01 --port-range 1/1/31-38
   ```

   For new clusters with HPE Aruba VSX switches, the `--mgmtsw` parameter requires you to specify only one of the two switches. The command detects and configures both HPE Aruba VSX switches in the partnership.

4. (Optional) Review the contents of the `cmcdetectd` file.

   For example:

   ```
   # cat /etc/cmc-switch-info.txt
   mac_address=00:03:e8:00:00:00, mgmtsw=sw-cdu01, vlans=3000, default_vlan=2000, bonding=manual,
   ports=1/1/1, redundant=yes, cmc_type=cmm, cmc_hostname=x1000c0
   mac_address=00:03:e8:01:00:00, mgmtsw=sw-cdu01, vlans=3000, default_vlan=2000, bonding=manual,
   ports=1/1/2, redundant=yes, cmc_type=cmm, cmc_hostname=x1000c1
   mac_address=00:03:e8:02:00:00, mgmtsw=sw-cdu01, vlans=3000, default_vlan=2000, bonding=manual,
   ports=1/1/3, redundant=yes, cmc_type=cmm, cmc_hostname=x1000c2
   mac_address=00:03:e8:03:00:00, mgmtsw=sw-cdu01, vlans=3000, default_vlan=2000, bonding=manual,
   ports=1/1/4, redundant=yes, cmc_type=cmm, cmc_hostname=x1000c3
   ...
   mac_address=00:03:e9:00:00:00, mgmtsw=sw-cdu01, vlans=3001, default_vlan=2001, bonding=manual,
   ports=1/1/11, redundant=yes, cmc_type=cmm, cmc_hostname=x1001c0
   ...
   # VLAN to management switch configuration
   vlan=3000, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostctrl
   vlan=3001, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostctrl
   vlan=3002, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostctrl
   vlan=3003, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostctrl
   vlan=2000, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostmgmt
   vlan=2001, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostmgmt
   vlan=2002, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostmgmt
   vlan=2003, mgmtsw=sw-cdu01, configured=no, chassis_type=cmm, vlan_type=hostmgmt
   ```

   **NOTE:** The preceding output was wrapped for inclusion in this documentation.

5. Use the `cmcdetectd` command in the following format to configure the switches and CMMs into the cluster manager database:

   ```
   cmcdetectd --switchconfig [--no-mgmt-config]
   ```

   Specify the `--no-mgmt-config` parameter if the switches are already configured in the cluster manager database. This would be the case, for example, for a reinstallation.

   Example 1. The following command configures `sw-cdu01`, configures `sw-cdu02`, adds all 32 CMMs into the database, and adds all 8 CECs into the database:

   ```
   # cmcdetectd --switchconfig
   ```

   Example 2. In the case of a reinstallation, you do not need to reconfigure the switches. The following command only adds the 32 CMMs and the 8 CECs into the cluster database:

   ```
   # cmcdetectd --switchconfig --no-mgmt-config
   ```

# Configuring the chassis controllers on an HPE Apollo 9000 cluster

**Procedure**

1. Enable the `cmcdetectd` service:

   ```
   # systemctl enable cmcdetectd
   ```

2. Configure the chassis controllers into the cluster:

   ```
   # systemctl start cmcdetectd
   ```

3. Wait for the chassis controllers to come up, and then verify that all chassis controllers appear in the output:

   ```
   # cm node show -t system chassis
   r1c1
   r1c2
   r1c3
   r1c4
   ```

# (Optional) Creating a chassis management controller (CMC) template file for an HPE Cray EX cluster

In an HPE Cray EX cluster, the `cmcinventory` service communicates with chassis management controllers (CMCs) to gather MAC address data from the compute nodes in each chassis slot. A CMC template file contains node-specific information about configuration attributes that you want to assign to the compute nodes in the cluster. When it runs, the `cmcinventory` service reads the CMC template file and assigns configuration attributes to the compute nodes.

For example, create a CMC template file if you want to assign specific hostnames, IP addresses, or other settings to compute nodes. You can specify configuration attributes based on node locations. You can specify information for only some nodes or for all nodes.

By default, the CMC inventory service configures the compute nodes with default configuration attribute values. For example, it assigns default hostnames to the compute nodes, and it assigns random IP addresses. If you want to use default configuration attribute values for all the compute nodes, you do not need to create a CMC template file.

**Procedure**

1. Enter the commands in the following table to retrieve general information about how to create the CMC template file:

| Command | Information returned |
|---|---|
| `cm node template help` | Overview and syntax information. |
| `man cluster-configfile` | Keywords used in the cluster definition file under the heading `DISCOVER SECTION`. |

2. Determine which IP addresses are already in use by the cluster.

   The CMC template file cannot assign existing, working IP addresses to compute nodes. The table in this step shows the commands that display IP addresses that are configured at this time. Do not specify these IP addresses in the CMC template file.

Enter the commands shown in the table, examine the output, and plan the IP address assignments you want to define. The IP address discovery commands are as follows:

| Command | Information returned |
|---|---|
| `cm network show` | Lists all available networks. Use if you want to assign IP addresses to nodes. |
| `cm network show -R -w` *network* | Returns subnet names and valid IP ranges. Make sure that the IP addresses you want to assign are within the valid IP range for the assigned network.<br><br>Specify a network name for *network*.<br><br>For example, the following command displays the valid range of IP addresses for *network*:<br><br>`cm network show -R -w` *network* |
| `cm node show -M` | Displays existing management network IP addresses. |
| `cm node show -B` | Displays existing iLO or baseboard management controller (BMC) IP addresses. |
| `cm controller show` | Displays the IP addresses for all existing controllers, including node controllers, switch controllers, and Gigabyte chassis management controllers (CMCs). |
| `cm group system show chassis` | Displays all existing chassis management modules (CMMs) |
| `cm node show -B -n` *CMM* | Displays IP addresses for a CMM. Specify a CMM hostname for *CMM*. |

3. Use a text editor to open a new file in which to specify CMC template information.

   This file can reside in any directory. For example:

   ```
   # cd /tmp
   # vi template
   ```

4. Edit the CMC template file to specify information for the compute nodes.

   You can use keywords to specify locations. The HPE Cray EX cluster location keywords are as follows:

| Location keyword | Characteristics |
|---|---|
| `rack_nr` | The `rack_nr` number begins with the configured starting rack number, which you can find with the following command:<br><br>`# cadmin –show-rack-start`<br>`1000` |
| `chassis` | Typically `0-7`. |
| `tray` | Typically `0-7`. This is sometimes referred to as a `slot` |
| `controller_nr` | Typically `0-1`, depending on the type of server. |
| `node_nr` | Typically `0-1`, depending on the type of server. |

Depending on what attributes you want to customize, use the location keywords as follows within the CMC template file:

- One line per server.

  Or

- Grouped into ranges. You can use brackets (`[ ]`) and hyphens (`–`) to specify ranges.

For example, the following lines use location keywords and ranges to describe one full rack of servers:

```
[discover]
# nodes in rack 1
rack_nr=1000, chassis=[0-7], tray=[0-7], controller_nr=[0-1],
node_nr=[0-1]
# controllers in rack 1
rack_nr=1000, chassis=[0-7], tray=[0-7], controller_nr=[0-1]
```

In the preceding, the third line defines all of the servers in the first rack, and the fifth line defines all of the controllers for those servers.

To see these ranges expanded, save and exit the text editor, and run the following command:

```
# cm node template show –c /tmp/template
rack_nr=1000, chassis=0, tray=0, controller_nr=0, node_nr=0,
rack_nr=1000, chassis=0, tray=0, controller_nr=0, node_nr=1,
rack_nr=1000, chassis=0, tray=0, controller_nr=1, node_nr=0,
rack_nr=1000, chassis=0, tray=0, controller_nr=1, node_nr=1,
rack_nr=1000, chassis=0, tray=1, controller_nr=0, node_nr=0,
rack_nr=1000, chassis=0, tray=1, controller_nr=0, node_nr=1,
rack_nr=1000, chassis=0, tray=1, controller_nr=1, node_nr=0,
rack_nr=1000, chassis=0, tray=1, controller_nr=1, node_nr=1,
rack_nr=1000, chassis=0, tray=2, controller_nr=0, node_nr=0,
rack_nr=1000, chassis=0, tray=2, controller_nr=0, node_nr=1,
rack_nr=1000, chassis=0, tray=2, controller_nr=1, node_nr=0,
.
.
.
```

5. Add your specifications to the basic hardware layout information.

For example, edit the topology to include more information, as follows:

```
# cat /tmp/template
[discover]
# nodes in rack 1
rack_nr=1000, chassis=[0-7], tray=0, controller_nr=[0-1], node_nr=0,
mgmt_net_name=hostmgmt2000, mgmt_net_ip=10.168.0.[101-116],
hostname1=blueoak[01-16]
# controllers for nodes in rack 1
rack_nr=1000, chassis=[0-7], tray=0, controller_nr=[0-1],
mgmt_bmc_net_name=hostctrl3000, mgmt_bmc_net_ip=10.176.0.[101-116],
node_controller
```

You cannot change the following values:

- The controller hostname of `hostname1`.
- The `mgmt_net_macs` configuration attribute and the `mgmt_bmc_net_macs` configuration attribute. The `cmcinventory` service detects the compute node MAC addresses.

**6.** (Optional) Add aliases for compute nodes.

To assign aliases to compute nodes within the CMC template file, include the predefined `cm-geo-name` alias group, and append your own alias definitions. The format to specify this alias group is as follows:

`alias_groups="cm-geo-name:x%Rc%Cs%Tb%Bn%N,`*`alias_group`*`:`*`alias_name`*`"`

The variables are as follows:

| Variable | Specification |
|---|---|
| *alias_group* | The name for the group of nodes. |
| *alias_name* | The name for the individual nodes. |

Note the following:

- There can be no space characters in the `alias_groups=` attribute specification.
- Enclose the attribute specification, which appears to the right of the equal sign (=), in quotation marks as shown.

Example. The `alias_groups=` attribute in the following line sets the hostname of each server in rack 1 to `n`*`X`* and creates `nidname` aliases for each node in `rack1`:

```
rack_nr=1000, chassis=[0-7], tray=[0-7], controller_nr=[0-1], node_nr=[0-1],
hostname1=n[1-128], alias_groups="cm-geo-name:x%Rc%Cs%Tb%Bn
%N,nidnames:nid[1-128]"
```

Step **8** includes an example of an `alias_groups` attribute.

For more information about assigning alias names to nodes, see the following:

- The following publication contains more information about assigning alias names to nodes:

  **HPE Performance Cluster Administration Guide**

- The following topic explains the `alias_groups` attribute:

  **`alias_groups`**

7. Save and close the file.

8. Verify the file.

   The verification command expands ranges in the file and displays individual assignments:

   For example:

   ```
   # cm node template show -c /tmp/template
   rack_nr=1000, chassis=0, tray=0, controller_nr=0, node_nr=0,
   hostname1=n1, alias_groups="cm-geo-name:x1000c0s0b0n0,nidnames:nid1"
   rack_nr=1000, chassis=0, tray=0, controller_nr=0, node_nr=1,
   hostname1=n2, alias_groups="cm-geo-name:x1000c0s0b0n1,nidnames:nid2"
   rack_nr=1000, chassis=0, tray=0, controller_nr=1, node_nr=0,
   hostname1=n3, alias_groups="cm-geo-name:x1000c0s0b1n0,nidnames:nid3"
   rack_nr=1000, chassis=0, tray=0, controller_nr=1, node_nr=1,
   hostname1=n4, alias_groups="cm-geo-name:x1000c0s0b1n1,nidnames:nid4"
   rack_nr=1000, chassis=0, tray=1, controller_nr=0, node_nr=0,
   hostname1=n5, alias_groups="cm-geo-name:x1000c0s1b0n0,nidnames:nid5"
   rack_nr=1000, chassis=0, tray=1, controller_nr=0, node_nr=1,
   hostname1=n6, alias_groups="cm-geo-name:x1000c0s1b0n1,nidnames:nid6"

   ...
   ```

9. Test the file.

   The `cm node template test` command confirms IP syntax and validates assigned IP addresses to ensure that they are within the subnet netmask.

   For example:

   ```
   # cm node template test -c /tmp/template
   ```

10. Submit the CMC template file to the `cmcinventory` service.

    For example:

    ```
    # cm node template submit -c /tmp/template
    ```

    The contents of the CMC template file are loaded into the cluster manager database and are applied to any new servers that the `cmcinventory` service discovers.

    To view the file contents at any time, enter the following command:

    ```
    # cm node template show
    ```

11. (Optional) Delete the CMC template contents from the cluster manager.

    After you submit the template file contents to the cluster manager, the contents remain in the cluster manager database for use whenever the `cmcinventory` service discovers new servers in locations that are not already populated in the cluster manager database.

    To delete the submitted CMC template contents and revert back to the `cmcinventory` default server settings, run the following command:

    ```
    # cm node template delete
    ```

    The command in this step prevents the `cmcinventory` service from accessing the CMC template file.

# Configuring the compute nodes into an HPE Cray EX cluster or an HPE Apollo 9000 cluster

**Procedure**

1. Open the following file in a text editor:

   ```
   /opt/clmgr/etc/cmcinventory.conf
   ```

2. Search for the line that appears as follows:

   ```
   # Enable/Disable individual services
   ```

   Edit the following fields to appear as follows:

   - On HPE Cray EX clusters:

     ```
     # Enable/Disable individual services
     cmc_inventory = False
     cmm_inventory = True
     ```

   - On HPE Apollo 9000 clusters:

     ```
     # Enable/Disable individual services
     cmc_inventory = True
     cmm_inventory = False
     ```

   - On HPE Cray EX clusters and on HPE Apollo 9000 clusters, peruse the file and set other fields as needed.

     For information about the settings in the `cmcinventory.conf` file, see the following:

     **Files used when configuring scalable unit (SU) leader nodes**

3. Save and close the `cmcinventory.conf` file.

4. Start the `cmcinventory` service to configure the compute nodes into the cluster.

   ```
   # systemctl enable cmcinventory
   # systemctl start cmcinventory
   ```

   **NOTE:** On HPE Cray EX clusters, if you created and submitted a CMC template file, the commands in this step apply that file to the compute nodes.

# Configuring the compute nodes into an HPE Apollo cluster with scalable unit (SU) leader nodes than is not an HPE Apollo 9000 cluster

If you have a cluster definition file that includes the following information about each compute node, use the procedure in this topic to configure the nodes into the cluster:

- The MAC addresses for the NICs in the compute nodes
- The MAC addresses for the node controllers in the compute nodes
- The compute node controller credentials

If you do not have a cluster definition file with the MAC addresses of the compute nodes, or if you do not have the MAC addresses at all, use the information in the following topic to configure the compute nodes into the cluster:

**Configuring compute nodes without a cluster definition file by using the `cm node discover` command**

**Procedure**

1. Verify that the cluster definition file for the compute nodes includes only information about the compute nodes.

   If the compute node cluster definition file includes information about power distribution units (PDUs) or compute nodes that you want to deploy as service nodes, save the lines for those additional components to another file. A later procedure explains how to configure those extra components into the cluster.

2. Verify that the file includes correct SU settings.

   For example:

```
# File compute.config
# Cluster definition file for compute nodes under 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=bt, 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=ib0, net_ifs="ib0"
template=su-compute, network=ib1, 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
```

3. Use the `cm node add` command to configure the compute nodes.

   `cm node add -c cluster_definition_file_for_compute_nodes`

   For *cluster_definition_file_for_compute_nodes*, specify the name of the compute node cluster definition file.

   For example:

   `# cm node add -c compute.config`

4. Create a compute node image, and deploy the image onto the compute nodes.

   Use the `cm node provision` command to provision the compute nodes.

   For example:

   `# cm node provision -n 'node*' -i image`

   For *image*, specify the name of the image file.

   For information about how to create a compute node image or the `cm node provision` command, see the following:

# Running the `discover` command

**NOTE:** The primary node discovery commands are the `cm node add` command and the `cm node discover add` command. The cluster manager supports the `discover` command as a secondary node discovery command.

The `discover` command performs the following actions:

- Adds component information to the cluster database and assigns component IP addresses and hostnames.
- Assigns images.
- Powers up nodes.
- Configures management switches.

Use the `discover` command in the following situations:

- During initial installation to configure all the components into the cluster.
- After initial installation to add or modify a component. The `discover` command can reconfigure an IP address, hostname, MAC address, bonding more, or other setting for a component.

**Procedure**

1. Use one of the following procedures to run the `discover` command:
   - **Running the `discover` command on an HPE Cray EX cluster or an HPE Apollo 9000 cluster with scalable unit (SU) leader nodes**
   - **Running the `discover` command on an HPE Apollo cluster with scalable unit (SU) leader nodes that is not an HPE Apollo 9000 cluster**

## Running the `discover` command on an HPE Cray EX cluster or an HPE Apollo 9000 cluster with scalable unit (SU) leader nodes

**Procedure**

1. Through an `ssh` connection, log into the admin node as the root user.
2. Run the `discover` command twice to configure the management switches and the SU leader nodes into the cluster.
   a. First, enter the `discover` command in the following format to update the cluster database with relevant templates:

      **`discover --configfile mgmtsw_suleader_file --update-templates`**

For *mgmtsw_suleader_file*, specify the cluster definition file that defines the management switches and the SU leader nodes.

**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 mgmtsw_suleader_file --all`**

For *mgmtsw_suleader_file*, specify the cluster definition file that defines the management switches and the SU leader nodes.

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
```

# Running the `discover` command on an HPE Apollo cluster with scalable unit (SU) leader nodes that is not an HPE Apollo 9000 cluster

**Procedure**

**1.** Verify the cluster definition files.

If you followed the procedure at the following link, you have two separate cluster definition files: One contains information about management switches. The other contains information about compute 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 to configure the management switches and SU leader nodes.

**a.** Enter the `discover` command in the following format to update the cluster database with relevant templates:

**`discover --configfile mgmtsw_suleader_file --update-templates`**

For *mgmtsw_suleader_file*, specify the name of the cluster definition file that defines the management switches and SU leader nodes.

**b.** 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 mgmtsw_suleader_file --all`**

For *mgmtsw_suleader_file*, specify the name of the cluster definition file that defines the management switches and SU leader nodes.

For example:

```
# discover --configfile mgmtsw_suleader.config --update-templates
# discover --configfile mgmtsw_suleader.config --all
```

**3.** (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
```

**4.** 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 are as follows:

| Variable | Specification |
|---|---|
| *hostname* | The hostname of the management switch |
| *new_password* | A strong, new password for the switch |

For example:

```
# switchconfig change_password --switches mgmtsw0 --new Hp3@dm!n2o20
```

**NOTE:** Hewlett Packard Enterprise strongly recommends that you implement standard and secure practices to store all passwords at your site. Do not lose this information.

**5.** Enter the following command to save the changed configuration to the nonvolatile memory (NVM) on the switches:

```
# switchconfig config -s all --save
```

**6.** (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
```

**7.** 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
```

# `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
- All other node types and component types

## `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 node controller 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 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 compute node

  The following command adds one 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 compute nodes, named `service1` through `service10`, to the cluster. The nodes have entries in the cluster definition file called `compute.config`. Within file `compute.config`, make sure that nodes 1 through 10 are listed sequentially. The command is as follows:

  ```
  # discover --configfile compute.config --nodeset 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
  ```

# (Conditional) Configuring cooling components

The HPE Cray EX clusters use the following cooling components:

- Cooling distribution units (CDUs), which are added to the cluster automatically and require no user installation or configuration actions.
- Rear-door heat exchangers.

The HPE Apollo clusters can use the following cooling components:

- HPE Adaptive Rack Cooling Systems (ARCS)
- Cooling distribution units (CDUs) (only on HPE Apollo 9000 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:

**Procedure**

1. **Configuring an HPE Cray EX rear door heat exchanger**
2. **Configuring an HPE Adaptive Rack Cooling System (ARCS) component**
3. **Configuring a cooling distribution unit (CDU) on an HPE Apollo 9000 cluster**

## Configuring an HPE Cray EX rear door heat exchanger

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 Administration Guide**

**Procedure**

1. Log in as the root user to the admin node.
2. Obtain the MAC address of the rear-door heat exchanger.

   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 rear-door heat exchanger.

   Use the `cm cooldev rdhx add` command in one of the following formats to enable the rear-door heat exchanger:

   - Format 1 - Adds the rear-door heat exchanger to the cluster based on its MAC address:

     ```
     cm cooldev rdhx add -m component_mac_addr -n hostname [-i ip_addr]
     ```

Use this command format the first time a rear-door heat exchanger is added to the cluster. This command requires you to provide the MAC address and a hostname.

- Format 2 - Adds the rear-door heat exchanger to the cluster using a previously assigned IP address:

  `cm cooldev rdhx add -n hostname -i ip_addr`

  Use this format, if the IP address was statically configured, is reachable, and is active on the rear-door heat exchanger.

The variables are as follows:

| Variable | Specification |
|---|---|
| *component_mac_addr* | The MAC address of the rear-door heat exchanger.<br><br>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 rear-door heat exchanger, or the hostname that is active on the rear-door heat exchanger. |
| *ip_addr* | In Format 1, you can specify an IP address, as follows:<br><br>• If you specify an IP address, make sure it is an active IP address. Such an IP address might have been assigned statically.<br>• If you do not specify an IP address, the cluster manager assigns an IP address, configures that IP address in DHCP, and enables the rear-door heat exchanger to obtain that IP address.<br><br>In Format 2, you do not specify the rear-door heat exchanger 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 rear-door heat exchanger to the cluster again after a maintenance period or outage. |

For more information about the commands to add, delete, or display rear-door heat exchangers, see the manpages for these commands or enter one or more of the following:

```
# cm cooldev rdhx -h
# cm cooldev rdhx add -h
# cm cooldev rdhx delete -h
# cm cooldev rdhx show -h
```

4. Repeat the preceding steps for each additional rear-door heat exchanger as needed.

# Configuring an HPE Adaptive Rack Cooling System (ARCS) component

After this procedure is complete, the ARCS component is enabled in the power and cooling infrastructure manager (PCIM). You can use PCIM to monitor the cooling components. For more information about PCIM, see the following:

**HPE Performance Cluster 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.<br><br>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:<br><br>• If you specify an IP address, make sure it is an active IP address. Such an IP address might have been assigned statically.<br>• If you do not specify an IP address, the cluster manager assigns an IP address, configures that IP address in DHCP, and enables the ARCS component to obtain that IP address.<br><br>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 ARCS component to the cluster again after a maintenance period or outage. |

For more information about the commands to add, delete, or display ARCS components, 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. Repeat the preceding steps for each additional ARCS component as needed.

# Configuring a cooling distribution unit (CDU) 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 Administration Guide**

**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.<br><br>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:<br><br>• If you specify an IP address, make sure it is an active IP address. Such an IP address might have been assigned statically.<br>• 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.<br><br>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. Repeat the preceding steps for each additional CDU as needed.

# 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` |
| 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
```

# (Conditional) Configuring power distribution units (PDUs) into the cluster

PDUs distribute AC power to the cluster components. PDUs are optional. The cluster manager requires you to configure the PDUs as a separate task. Use the information in this procedure to configure the PDUs into the cluster.

On HPE Cray EX clusters and HPE Apollo clusters the PDUs reside in each rack. For these clusters, and for all other clusters with PDUs that reside in racks, include the PDUs in the cluster definition file.

For example, assume that you need to include a definition for `pdu0`. A line such as the following in the cluster definition file configures 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"
```

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 Administration Guide**

**Procedure**

1. Use a text editor to create a file for the PDUs.

   For example, create file `pdu.config`.

   If you have a cluster definition file that includes PDU information, copy the PDU information from the cluster definition file into the PDU-specific file, and proceed to the following step:

   Step **4**

2. Include the following information in this file:

   - Specify the network upon which the PDU resides. In e example in this procedure case, it is `head-bmc`, which specifies the head BMC network.
   - You can specify a geographic location 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.

     The protocol is SNMP on HPE Cray EX and HPE Apollo clusters with scalable unit (SU) leader nodes.

When you specify the SNMP protocol, you can 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"
```

For example, create a file that includes information similar to the following:

```
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"
```

3. Save and close the file.
4. Use the `cm node add` command to configure the PDUs into the cluster.

   The format is as follows:

   ```
   cm node add -c cluster_definition_file_for_PDUs
   ```

   For *cluster_definition_file_for_PDUs*, specify the name of your cluster definition file.

   For example:

   ```
   # cm node add -c pdu.config
   ```

# Configuring compute nodes that are not under the control of a leader node

Use the commands in this chapter to add compute nodes that do not reside in a chassis. These might be extra compute nodes deployed with user services. If a compute node resides in a chassis, use the `cmcinventory` command to add them to the cluster.

You can use the procedures in this chapter later if you add nodes or components to the cluster.

**Procedure**

1. Enter the following command, examine the output, and verify that all compute nodes have been added to the cluster:

   ```
   # cm node show
   ```

   If a compute node resides in a chassis, it should appear in the command output. If a node that resides in a chassis does not appear in the command output use the `cmcinventory` service to add the node into the cluster.

   If a compute node does not appear in the command output because it is not yet configured into the cluster, continue with this procedure. This is the case for nodes that are not under the control of a leader node. For example, this is the case for compute nodes deployed as login nodes.

2. Use one or both of the following procedures to configure compute nodes into the cluster:

   - **Configuring compute nodes with a cluster definition file and the `cm node add` command**. Use this command if you have a cluster definition file that includes the compute nodes.
   - **Configuring compute nodes without a cluster definition file by using the `cm node discover` command**. Use this procedure if you do not have a cluster definition file that includes the compute nodes.

## Configuring compute nodes with a cluster definition file and the `cm node add` command

The `cm node add` command adds components, such as compute nodes or racks of multiple compute nodes, to a cluster. You can use this command to add many types of cluster components, but this topic specifically addresses compute nodes.

The command in this topic assumes that you have a cluster definition file that includes the following information for each compute node:

- The MAC address for the NIC
- The MAC address for the node controller
- The node controller credentials

For more information about the parameters to this command, enter the following:

```
cm node add -h
```

**Procedure**

1. Obtain or create a cluster definition file that includes compute node information.

Include configuration attributes for the MAC addresses, IP addresses, and other information.

For example, assume that `computes.config` is a cluster definition file with the following contents:

```
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,card_type=iLO,\
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
```

2. Enter the following command:

   `cm node add -c` *`cluster_definition_file_for_new_nodes`*

   For *cluster_definition_file_for_new_nodes*, specify the name of your cluster definition file.

   For example:

   ```
   # cm node add -c computes.config
   ```

3. Use the `cm node provision` command to provision the new compute nodes with an image and (optionally) to power cycle the new compute nodes.

# Configuring compute nodes without a cluster definition file by using the `cm node discover` command

The `cm node discover` command can configure scalable unit (SU) leader nodes and compute nodes into the cluster without the use of a cluster definition file.

This command assumes the following:

- You do not have a cluster definition file that includes the nodes you want to add.
- The SU leader nodes or the compute nodes are capable of being PXE booted.
- For the nodes you want to add, you do not know the MAC addresses of the node controllers or the MAC addresses of the NICs . If you know the MAC address information for the nodes you want to add, use the `cm node add` command to add the node.

Whether you have the MAC addresses or not, you can use `cm node discover` to set the node controller credentials. This command PXE boots a small operating system on the node to gather node information and (optionally) set credentials.

The `cm node discover` command guides you through an automated, incremental process for building a cluster definition file for adding new nodes to the cluster.

For more information about the parameters to this command, enter the following:

```
# cm node discover -h
```

To display help for the steps in this process, enter the following command:

```
# cm node discover help
```

**Procedure**

1. Verify that the new SU leader nodes or the new compute nodes are cabled and plugged in.
2. Log into the admin node as the root user.
3. Enter the following command to create a pool of IP addresses, with a short lease time, in the DHCP service:

   ```
   # cm node discover enable
   ```

If necessary, specify additional parameters. For example, you can specify the following:

- A specific subnet for the pool of IP addresses.
- A specific miniroot for operating system discovery.

4. Manually press the power-on button for each of the new SU leader nodes or compute nodes.

   As each SU leader node or compute node powers up, the cluster manager grants a leased IP address from the pool, and the miniroot environment boots.

5. Enter the following command and observe the leased IP address information:

   ```
   # cm node discover status
   ```

   This command lists all the leased IP addresses and uses `ssh` to connect to each of these leased IP addresses. The command is trying to detect whether the nodes have PXE booted the cluster manager miniroot operating system. When the `ssh` attempt is successful, the cluster is in contact with the new compute node.

6. Make sure that the `cm node discover status` command shows all the nodes you want to add.

   Do not proceed to the next step until all nodes are shown in the output.

7. Enter the `cm node discover mkconfig` command, in a format similar to the following, to generate a cluster definition file for the new nodes:

   ```
   cm node discover mkconfig -o "bmc_username=uname, bmc_password=pwd"
   cluster_definition_file
   ```

   The variables are as follows:

| Variable | Specification |
|---|---|
| *uname* | The BMC username you want to assign to the node controllers. |
| *pwd* | The BMC password you want to assign to the node controllers. |
| *cluster_definition_file* | The name for the output file, which becomes the cluster definition file for these nodes.<br><br>For example, `suleader.config` or `computes.config`. |

   The BMC credentials are required. This command creates a cluster definition file with very minimal entries for each new node. To add other common settings per node, expand the content in the `-o` option. For example, to configure the console to be `ttyS1`, change the `-o` option to the following:

   ```
   -o “bmc_username=username, bmc_password=password, console_device=ttyS1”
   ```

   For more information about the settings you can include on the `-o` option, see the following:

   **Specifying configuration attributes**

8. (Optional) Add node-specific settings in the cluster definition file.

   At this point, you have a cluster definition file. If you want to specify node-specific settings, edit the cluster definition file now.

9. Enter the `cm node discover add` command, in the following format, to add the new compute node to the cluster manager database:

   ```
   cm node discover add [-s] [-i image] [-d disk] cluster_definition_file
   ```

   This command adds the new nodes and resets the node controllers so that they pick up appropriately configured IP addresses.

The parameters and variables are as follows:

| Parameter or variable | Specification |
|---|---|
| `-s` | Specify the `-s` parameter if the BMC credentials in the cluster definition file need to be configured in the BMC.<br><br>If the BMC credentials are not configured in the BMC, this option is not needed. |
| `-i` *`image`* | The image you want to assign to the compute nodes.<br><br>If you specify an image, the command reboots the nodes and provisions the nodes with the specified image. Otherwise, by default, this command powers off the nodes, which postpones provisioning.<br><br>If you do not specify the `-i` option, the cluster manager powers down the nodes. You can use the `cm node provision` command to deploy an image to the nodes. |
| `-d` *`disk`* | Specify the `-d` *`disk`* parameter if you also specify the `-i` *`image`* parameter.<br><br>For *disk*, specify the disk to install the *image*. The default is `/dev/sda`. |
| *cluster_definition_file* | The name of the cluster definition file for these nodes, which you created in the following step:<br><br>Step **7** |

**10.** Enter the following command to delete the pool of IP addresses from the DHCP service:

```
# cm node discover disable
```

# (Conditional) Adding controllers manually

The cluster manager adds most types of controllers to the cluster database automatically. However, the cluster manager does not add the following controllers or components to the database automatically:

- An external HPE Slingshot switch controller
- A Gigabyte chassis controller

As a troubleshooting tactic, you can also use the `cm controller` command to delete and then to add a misconfigured controller. Use `cm controller delete` to delete the misconfigured controller and then `cm controller add` to add the controller back in correctly.

If your cluster contains any of the preceding controller types, complete the procedure in this topic to add the controllers manually.

**Procedure**

1. Use the `cm controller add` command to configure the controller into the cluster database.

   The format of this command is as follows:

   `cm controller add -c` *`hostname`* `-t` *`controller_type`* `-m` *`mac_address`* `-u` *`username`* `-p` *`password`*

   The variables are as follows:

   | Variable | Specification |
   |---|---|
   | *hostname* | The hostname you want to assign to the controller. |
   | *controller_type* | Enter one of the following keywords depending on the type of controller you want to add:<br>• `external_switch`. Use this keyword for an external HPE Slingshot switch controller.<br>• `gigabyte`. Use this keyword for a Gigabyte chassis controller. |
   | *mac_address* | The MAC address of the controller. |
   | *username* | The username used to log into the controller. |
   | *password* | The password used to log into the controller. |

2. Enter the `cm controller show` command to display the information for the controller you just added.

   The format for this command is as follows:

   `cm controller show -c` *`hostname`*

   For *hostname*, enter the hostname of the controller you just added.

   For example:

   ```
   # cm controller show -c x9000c1r3b0
   NAME        TYPE                  ADMINISTRATIVESTATUS PROTOCOL CHANNEL MACADDRESS        IPADDRESS  IPV6ADDRESS
   x9000c1r3b0 cmm_switch_controller online               None     None    XX:XX:XX:XX:XX:XX XX.XXX.X.X None
   ```

3. Repeat the preceding steps to configure all controllers into the cluster.

**NOTE:** If you have many controllers, you can create a file with controller information and specify that file as an argument to the following command:

```
cm node add -c input_file
```

This single command adds multiple controllers. For more information, enter the following command:

```
# cm node add -h
```

## Using the `cm controller add` command

The `cm controller add` command adds an external switch controller or a Gigabyte controller to the cluster database. For more information, enter the following command:

```
# cm controller add -h
```

## Using the `cm controller show` command

The `cm controller show` command displays information for all controllers of all types.

If you enter the command without any arguments, it displays all the controllers in the cluster. For example::

```
# cm controller show

NAME        TYPE                          ADMINISTRATIVESTATUS PROTOCOL                  CHANNEL MACADDRESS
x9000c1r3b0 cmm_switch_controller         online               None                      None    XX:XX:XX:XX:XX:XX
x9000c1r7b0 cmm_switch_controller         online               None                      None    XX:XX:XX:XX:XX:XX
x9000c1s0b0 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s0b1 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s1b0 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s1b1 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s2b0 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s2b1 cmm_node_controller           online               None                      None    XX:XX:XX:XX:XX:XX
x9000c1s3b0 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c1s3b1 cmm_node_controller           online               Cray,NO_IPMI,None,redfish None    XX:XX:XX:XX:XX:XX
x9000c3r3b0 cmm_switch_controller         online               None                      None    XX:XX:XX:XX:XX:XX
x9000c3r7b0 cmm_switch_controller         online               None                      None    XX:XX:XX:XX:XX:XX
x9000cec0   cabinet_environment_controller online               None                      None    None
x9000cec1   cabinet_environment_controller online               None                      None    None
```

**NOTE:** The preceding output was truncated from the right for inclusion in this documentation.

## Using the `cm controller delete` command

The `cm controller delete` command deletes a controller from the cluster database. For more information about this command, enter the following:

```
# cm controller delete -h
```

# Backing up the cluster

**Procedure**

1. **Backing up the admin node**
2. **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.

**Prerequisites**

**Backing up the admin node**

**Procedure**

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

   If the cluster has a high availability (HA) admin node, log into the virtual machine (VM) admin node.

2. 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 the `cm node add` command, the `cm node discover add` command, 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.

3. Copy the cluster definition file to another server at your site.

4. (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`

5. (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`

6. Enter the following commands to stop the cluster manager:

```
# systemctl stop config_manager.service
# systemctl stop clmgr-power.service
# systemctl stop cmdb.service
```

7. Enter the following command to back up the cluster database:

```
# 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"
```

8. Enter the following commands to start the cluster manager:

```
# systemctl start cmdb.service
# systemctl start clmgr-power.service
# systemctl start config_manager.service
```

---

**NOTE:** In the future, to restore the cluster database and start the cluster manager, see the procedure in the following:

**HPE Performance Cluster Administration Guide**

---

9. 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.

10. Enter the following command to save the changed configuration to the nonvolatile memory (NVM) on the switches:

```
# switchconfig config -s all --save
```

11. 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
```

12. 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, complete the following procedure:

  **Naming the storage controllers for clusters with a system admin controller high availability (SAC HA) admin node**

- 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 nodes

You can use post-installation scripts to customize operations on compute nodes and on scalable unit (SU) 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 compute nodes into network groups. It is most common to configure all the compute nodes under a common switch into a network group.

For information about how to configure network groups, see the following:

**HPE Performance Cluster Administration Guide**

# Naming the storage controllers for clusters with a system admin controller high availability (SAC HA) admin node

Complete the procedure in this topic if the cluster has a SAC 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. Log into the admin node as the root user.
2. 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. |

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 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:

- HPE Slingshot fabric IP addresses. For example, `hsn0.cm.clusterdomain.com` or `hsn1.cm.clusterdomain.com`.
- InfiniBand fabric IP addresses. For example, `ib0.cm.clusterdomain.com` or `ib1.cm.clusterdomain.com`.
- Management fabric IP addresses. For example, `head.cm.clusterdomain.com`, `hostmgmt.cm.clusterdomain.com`, or `gbe.cm.clusterdomain.com`.
- Public or external IP addresses. For example, `cm.clusterdomain.com` or `public.clusterdomain.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 DNS search order**
- **Retrieving the DNS search order**

## Analyzing your environment

Sometimes a host includes multiple network interfaces.

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.cm.clusterdomain.com cm.clusterdomain.com`

    or
  - `ib0.cm.clusterdomain.com gbe.cm.clusterdomain.com cm.clusterdomain.com`
- The host lookup command returns the `gbe` IP address when the search path is one of the following:
  - `gbe.cm.clusterdomain.com cm.clusterdomain.com`

    or
  - `gbe.cm.clusterdomain.com ib0.cm.clusterdomain.com cm.clusterdomain.com`
- 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.

The following topics explain how to view or configure the global or per-node search order:

- **Configuring the DNS search order**
- **Retrieving the DNS search order**

## Configuring the 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] [-n` *`node`*`] --domain-search-path` *`new_domain_search_path`*

   The variables are as follows:

| Variable or parameter | Specification |
|---|---|
| `-g` | Conditional. Use when you want to configure the global search order. |
| *node* | Conditional. Use when you want to configure the search order for one node. Specify the node hostname. |
| *new_domain_search_path* | 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.clusterdomain.com.com,head.cm.clusterdomian.com
```

Example 2. The following command sets the domain search path for `n0`:

```
admin:~ # cm node set -n n0 --domain-search-path head.cm.clusterdomain.com,ib0.cm.clusterdomain.com
```

## Retrieving the 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:

   ```
   cm node show --domain-search-path [-n node]
   ```

   For *node*, specify a node hostname. Specify this optional 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:

   ```
   # cm node show -g --domain-search-path
   ib0.cm.clusterdomain.com,head.cm.clusterdomain.com
   ```

   Example 2. The following command retrieves the domain search path for one node, `n0`:

   ```
   # cm node show --domain-search-path -n n0
   head.cm.clusterdomain.com,ib0.cm.clusterdomain.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 compute node as a secondary DNS server when the admin node is unavailable. You can configure a backup DNS only after the cluster is configured completely. This feature is optional.

The following procedure explains how to configure a 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 compute nodes:

   ```
   # cnodes --compute
   ```

The preceding command lists all nodes that are classified as compute nodes, so the list includes scalable unit (SU) leader nodes and fabric management nodes. Select a compute node for use as the backup DNS. Do not select an SU leader node or a fabric management node for the backup DNS.

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

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

4. On the **Main Menu** screen, select **D Configure Domain Name System (DNS)**, and select **OK**.
5. On the **Domain Name System (DNS) Menu** screen, select **B Configure Backup DNS Server (optional)**, and select **OK**.
6. On screen that appears, enter the identifier for the compute node that you want to designate as the backup DNS, and select **OK**.

   For example, you could configure 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.

# Setting a static IP address for the node controller 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 node controller.
- You want to configure a high availability (HA) admin node. In this case, perform this procedure on the node controllers on each of the two admin nodes.

When you set the IP address for the node controller 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 node controller 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 node controller is a static IP address.
- The command sets the IP address to the IP address that is currently assigned to the node controller.

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 node controller. |
| *netmask* | The netmask you want to assign to the node controller. |
| *gateway* | The gateway you want to assign to the node controller. |

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 Array Services for HPE Message Passing Interface (MPI) programs

You can configure compute nodes into an array. After you configure a set of 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**

You cannot include the admin node or any leader nodes 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 compute nodes configured for user services.

After you create the images, you can push out the new images.

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.

**Procedure**

1. **Planning the configuration**
2. **Preparing the Array Services images**
3. Complete one of the following:
   - **(Conditional) Permitting remote access to the service node**

     Or

   - **(Conditional) Preventing remote access to the service node**
4. **Distributing images to all the nodes in the array**
5. **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 display information about compute nodes, enter the following command:

   ```
   # cm node show -t system compute
   ```

   The command output includes information about nodes that might be configured as service 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 scalable unit (SU) leader nodes that is running in production mode:

   ```
   # cm image show
   su-sles15spX                # original, factory-shipped system image
   su-sles15spX.prod1          # customized image for this cluster
   sles15spX                   # original, factory-shipped system image
   sles15spX.prod1             # customized image for this cluster
   ```

   The output includes image `sles15spX.prod1`. The `sles15spX.prod1` image is installed on a compute node that is configured as a service node. Image `sles15spX.prod1` is based on image `sles15spX`, 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`

   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 service nodes and `none` on the compute nodes

     or

     `noremote` on the service 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 service nodes:
  - `noremote` on the service 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.

## 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 service node

     And
   - One of the images that resides on a compute node

   The format is as follows:

   `cm image copy -o existing_image -i new_image`

   The variables are as follows:

| Variable | Specification |
|---|---|
| *existing_image* | The name of one of the existing images. |
| *new_image* | The new name for that to want to give to the image. |

   For example, the following command copies the first-generation compute node production image to a new, second-generation production image:

   `# cm image copy -o sles15spX.prod1 -i sles15spX.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 service 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_service_image/etc/munge/munge.key \
/opt/clmgr/image/images/new_compute_image/etc/munge/munge.key
```

The variables are as follows:

| Variable | Specification |
|---|---|
| *new_service_image* | The name of the new service node image you created. |
| *new_compute_image* | The name of the new compute node image you created. |

For example:

```
# cp /opt/clmgr/image/images/sles15spX.prod2/etc/munge/munge.key \
/opt/clmgr/image/images/ice-sles15spX.prod2/etc/munge/munge.key
```

5. Use the following command to install the new image on the service node:

   ```
   cm node provision -n hostname(s) -i new_service_image -s
   ```

   The variables are as follows:

   | Variable | Specification |
   |---|---|
   | *hostname(s)* | The hostname or hostnames of the service node. This node is the node that you want users to log into when they log into the array. |
   | *new_service_image* | 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 sles15spX.prod2 -s
   ```

6. Use the `ssh` command to log into the service 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 service node and compute nodes into an array.

   You can specify more than one service 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 nodes ...
   ```

   The variables are as follows:

| Variable | Specification |
|---|---|
| *arrayname* | A name for the array. The default is `default`. |
| *method* | `munge`, `none`, or `simple`. A later step explains how to specify `noremote` for a service node. |
| *nodes* | A list of node IDs. |

## (Conditional) Permitting remote access to the service node

Complete this procedure in the following circumstances:

- 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 service node remotely to submit MPI from HPE programs. The service node is assumed to be a compute node.

The following procedure assumes that you want to permit job queries and commands on the service node. It explains how to copy the array daemon files to the admin node.

**Procedure**

1. Log into one of the service nodes as the root user.
2. Copy the `arrayd.auth` file and the `arrayd.conf` files from the service node to the new service node image on the admin node.

   Enter the following command:

   ```
   # scp /etc/array/arrayd.* \
   admin:/opt/clmgr/image/images/service_image/etc/arrayd.*
   ```

   For *service_image*, specify the service 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/sles15spX.prod2/etc/arrayd.*
   ```

3. Copy the `arrayd.auth` file and the `arrayd.conf` files from the service 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. This is a compute node image.

   Enter this command all on one line.

## (Conditional) Preventing remote access to the service node

Complete the procedure in this topic if you specified `-A none` for authentication, and you want to prevent users from logging into a service node remotely to submit MPI from HPE programs

This procedure explains how to prevent a service node from receiving any requests from other computers on the network. In this case, the service 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 service 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 service node to the new service node image on the admin node:

   **`# scp /etc/array/arrayd.* \`**
   **`admin:/opt/clmgr/image/images/`*`service_image`*`/etc/arrayd.*`**

   For example:

   **`# scp /etc/array/arrayd.* \`**
   **`admin:/opt/clmgr/image/images/sles15spX.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/sles15spX.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`**
11. Use the `cm image revision commit` command to back up the images.

## Distributing images to all the nodes in the array

The following procedure explains how to complete the following tasks:

- Assign the new service node image to the service 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 service node image to the service nodes.

   Use one or more `cm node provision` commands in the following format:

   `cm node provision -n` *`hostname`* `-i` *`new_service_image`* `-s`

   The variables are as follows:

| Variable | Specification |
|---|---|
| *hostname* | Specify the hostname for one or more of the service nodes. In the cluster definition file, this name appears in the `hostname1` field.<br><br>You can specify the `*` wildcard character to represent a string of identical characters in this field. Use wildcard characters in the following situation:<br><br>• If you want to specify more than one hostname<br><br>and<br><br>• If your nodes have names that are similar<br><br>For example, if your hostnames are `n1`, `n2`, `n3`, and `n57`, specify `n*` in this field if you want to specify all service nodes. |
| *new_service_image* | Specify the name of the new service node image you created. |

   Example 1. The following command assigns the new service node image to all service nodes:

   ```
   # cm node provision -n n* -i sles15spX.prod2 -s
   ```

   Example 2. The following command assigns the new service node image to `n101`:

   ```
   # cm node provision -n n101 -i sles15spX.prod2 -s
   ```

## Power cycling the nodes and pushing out the new images

### Procedure

1. Enter the following command to reboot the service 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.

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 compute nodes that use an NFS root file system.

   On clusters with scalable unit (SU) leader nodes, the commands in this step can take a few moments to complete.

   The following tables contain the instructions you need to complete this step.

   **Complete the following steps on clusters with compute nodes that have NFS root file systems:**

| Step | Task |
|---|---|
| a. | Run the `cm image activate` command in the following format to activate the NFS image:<br><br>`cm image activate -i new_compute_image`<br><br>For *new_compute_image*, specify the name of the new compute node image you created.<br><br>For example:<br><br>`# cm image activate -i sles15spX.prod2`<br>`Activate image - Syncing image to RO NFS path`<br>`.`<br>`.`<br>`.` |
| b. | Assign the new compute node images to the compute nodes.<br><br>For example, use the following `cimage` command:<br><br>`# cimage --set sles15spX.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.

# Creating a ComputeNode image for a node running the RHEL 7 operating system

**Procedure**

1. Use the following commands to register the RHEL 7 ComputeNode repository and RHEL 7 cluster manager repository:

   ```
   cm repo add rhel_iso_file
   cm repo add cluster_manager_iso_file
   ```

   For *rhel_iso_file* and *cluster_manager_iso_file*, specify the .iso file names.

   For example:

   ```
   $ cm repo add RHEL-7.9-20200225.1-ComputeNode-x86_64-dvd1.iso
   $ cm repo add cm-1.6-cd1-media-rhel79-x86_64.iso
   ```

2. Create a symbolic link for the repositories, and add the new repository as a custom repository.

   For example:

   ```
   $ ln -s /opt/clmgr/repos/cm/Cluster-Manager-1.6-rhel79-x86_64/rhel7cn/
   /opt/clmgr/repos/rhel7cn
   $ cm repo add /opt/clmgr/repos/rhel7cn --custom Rhel7.9_ComputeNode_Extras
   ```

3. Create a repository group to use when creating images.

   For example:

   ```
   $ cm repo group add cm_rhel79_compute --repos
   Red-Hat-Enterprise-Linux-ComputeNode-7.9-x86_64
   Cluster-Manager-1.6-rhel79-x86_64 Rhel7.9_ComputeNode_Extras
   ```

4. Create a RHEL 7 compute image using the repository group.

   For example:

   ```
   $ cm image create -i rhel7.9_compute --repo-group cm_rhel79_compute \
   --rpmlist /opt/clmgr/image/rpmlists/generated/generated-group-cm_rhel79_compute.rpmlist
   ```

# Creating security certificates from a site-specific certificate authority (CA)

The cluster manager includes a security certificate. By default, the REST API and web server use the security certificate that the cluster manager provides.

If your site has a trusted security certificate from a CA, you can complete the procedure in this topic to regenerate the cluster security certificates based on your site certificate.

The procedure in this topic does not change the certificates in the secrets bundle on leader nodes.

**Procedure**

1. Log into the admin node as the root user.
2. Use a text editor to open the following file:

   ```
   /opt/clmgr/tools/gen-custom-rest-certs
   ```

3. Edit the following fields in the file:

| Field | Specification for this cluster |
|---|---|
| caCert | The full path to the CA certificate |
| serverKey | The full path to the server key |
| serverCert | The full path to the server certification |

4. Save and close file gen-custom-rest-certs.
5. Enter the following command to create new security certificates:

   ```
   # /opt/clmgr/tools/gen-custom-rest-certs
   ```

6. Change to the custom-certs directory:

   ```
   # cd custom-certs
   ```

7. Enter the following commands to copy the newly generated certificates, the Java keystore, and the private keys to the appropriate directory:

   ```
   # cp *.pem /opt/sgi/secrets/CA/cert/
   # cp *.p12 /opt/sgi/secrets/CA/private/
   # cp *.key /opt/sgi/secrets/CA/private/
   ```

8. Use a text editor to open the following file:

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

9. Locate the line that begins with CMU_JAVA_SERVER_ARGS, and add the following string within the quotation marks (" "):

   ```
   -Djdk.security.allowNonCaAnchor=true
   ```

   For example, after adding the new string, the line might look as follows:

   ```
   CMU_JAVA_SERVER_ARGS="-Djdk.security.allowNonCaAnchor=true"
   ```

   If other strings reside within the quotation marks, put this new string at the end.

10. Save and close the cmuserver.conf file.
11. Restart the cluster manager:

    ```
    # systemctl restart cmdb.service
    # systemctl restart clmgr-power.service
    # systemctl restart config_manager.service
    ```

12. Create a tar file for the new security certificates:

    ```
    # tar cJf \
    - root-ssh/compute udpcast CA/private/user_client.key \
    CA/private/admin_client.key CA/cert \
    | openssl enc \
    -md md5 -aes-256-cbc -pass file:/opt/sgi/secrets/bootstrap/passwd.txt \
    -out /opt/sgi/secrets/bootstrap-secrets/compute.tar.xz.aes
    ```

13. For compute nodes that are not under the control of a leader node, use the cm node provision command to mark compute nodes for reimaging.

    The compute nodes acquire the new secrets, including the certificates, when they are reprovisioned.

For information, see the following:

**HPE Performance Cluster Administration Guide**

# 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 service nodes with the new, reconfigured image.
- Edit the node customization scripts.

## 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.

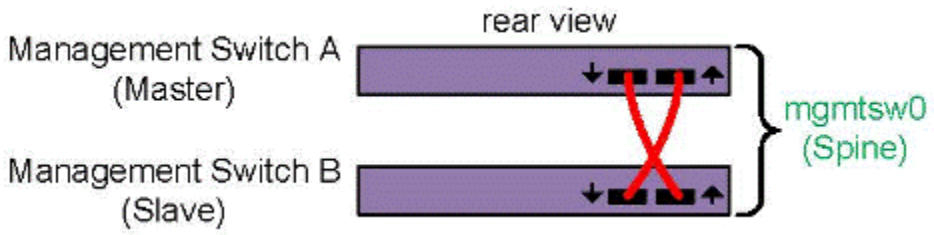


**Figure 2: Spine switch stack with two switches**

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.

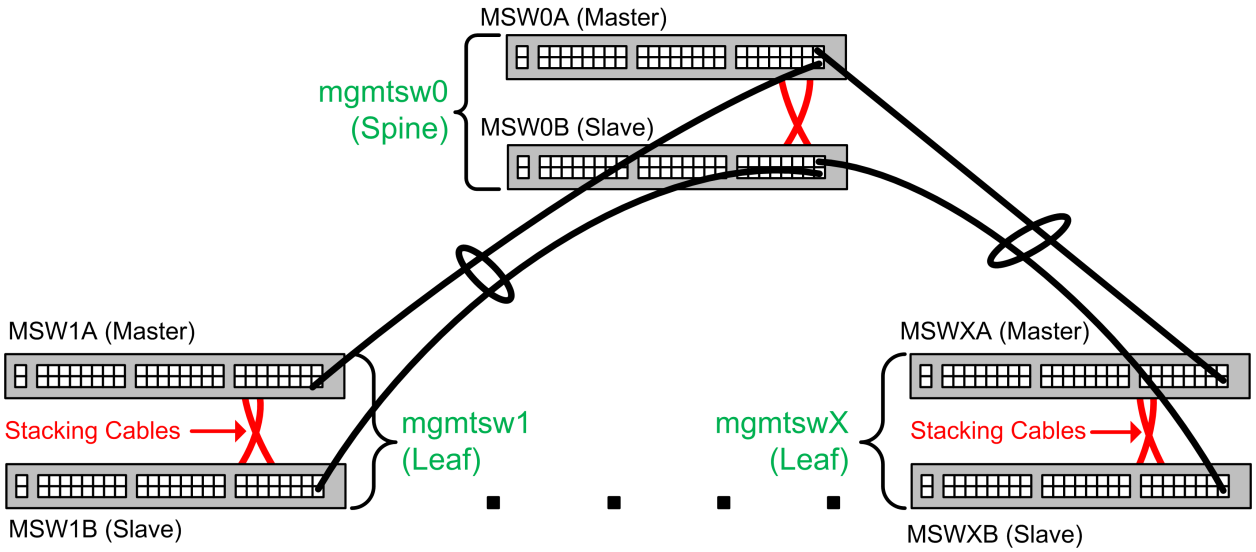


**Figure 3: Switch stack with multiple switches**

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 node discovery commands. During the node discovery 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 other than the admin node must begin in a powered off state. Console into the switch, and configure the switch manually.

3. (Conditional) Disconnect the secondary, redundant cables that connect switches together.

   Complete this step if you have not yet configured the management switches into the cluster. 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 scalable unit (SU) leader nodes 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 leader1
# switchconfig_configure_node --node 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 controllers.

  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.

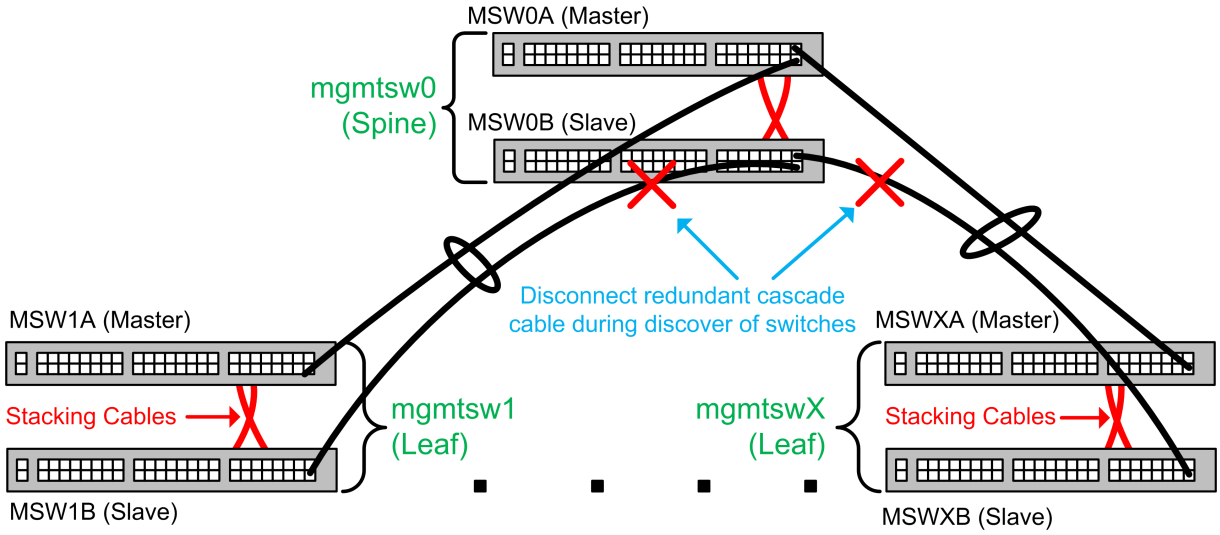


**Figure 4: Cables to disconnect**

# `cmcinventory` service fails to copy `ssh` keys to the controllers

### Symptom

A problem can occur on HPE Cray EX platforms when you attempt to `ssh` to a node or to open a console, and the system prompts you to enter a password.

### Cause

When the cluster configuration process completes as expected, the installer copies `ssh` keys to the following hardware controllers:

- `nC` (node controllers)
- `cC` (chassis controllers)

  If the automatic distribution process fails, the `ssh` keys are not copied to the controllers as expected. Use the `cbios` command to distribute the `ssh` keys manually.

For information about the `cbios` command, enter the following:

```
# cbios -h
```

### Action

1. Log into the admin node as the root user.
2. Enter one of the following commands to add the `ssh` keys to the node controller:

   - If the system displays an `ssh` password prompt when you try to log into a controller, enter the following command:

     ```
     cbios --AdminAdd -n hostname
     ```

For *hostname*, enter the hostname of the controller that returns an `ssh` password prompt. You can enter one or more controller hostnames. Use a comma to separate multiple hostnames.

- If the node prompts for a password when you try to open a console to a node, enter the following command:

  ```
  cbios --ConsoleAdd -n hostname
  ```

3. Enter the following command to verify the effect of the `cbios` command:

   ```
   cbios --Show -n hostname
   ```

# Chassis controllers failed to configure

If you suspect that the chassis controllers failed to configure automatically, look in the following log file for information regarding the status of chassis controllers in the system:

```
/var/log/cmcdetectd.log
```

If you find errors in the preceding log files, power on the chassis controllers and complete the following steps to configure the chassis controllers:

1. **Reviewing the chassis controller configuration**
2. Complete one of the following procedures:

   - **Method 1 - Configuring the chassis controller switches manually**
   - **Method 2 - Configuring the chassis controller switches manually**

## `cmcdetectd` daemon

The `cmcdetectd` daemon runs on the admin node. This daemon uses a specific `tcpdump` command to listen to DHCP packets generated by the chassis controller on the management network. When the `cmcdetectd` daemon receives a DHCP packet generated by the chassis controller, 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 chassis controller in question.

Chassis controllers are cabled in the same manner as other redundantly cabled components in the cluster. For example, assume that rack 1, chassis controller 1, port 1 is connected to `mgmtsw0` port 1/0/11. In this case, rack 1, chassis controller 1, port 2 must be connected to `mgmtsw0` port 2/0/11, and so on.

## `cmcdetectd` daemon on the HPE Cray EX cluster

Unlike the HPE Apollo 9000 system management, HPE Cray EX system management uses tagged VLANs to isolate the types of traffic on compute nodes. This practice comes with some noted differences in the cluster manager and in how `cmcdetectd` is invoked. As a general rule, the cluster manager recommends that VLAN numbers are dictated in the following logic structure:

- Untagged (native) VLAN = *cabinet_number* + 1000
- Tagged VLAN = *cabinet_number* + 2000

For example, assume that a cluster has four HPE Cray EX cabinets numbered x1000 through x1003. The VLANs associated with those cabinets are as follows:

- x1000 = 2000 (untagged) & 3000 (tagged)
- x1001 = 2001 (untagged) & 3001 (tagged)
- x1002 = 2002 (untagged) & 3002 (tagged)
- x1003 = 2003 (untagged) & 3003 (tagged)

Before `cmcdetectd` is run, use one of the following methods to set the following attributes so that they match these numbers:

- Edit the cluster definition file and add information in the `[attributes]` section. When you run the `configure-cluster` command, the cluster manager sets these values.

  The following is an example HPE Cray EX cluster definition file `[attributes]` section:

  ```
  [attributes]
  mgmt_net_routing_protocol=ospf
  rack_start_number=1000
  mgmt_vlan_start=2000
  mgmt_vlan_end=2999
  mgmt_ctrl_vlan_start=3000
  mgmt_ctrl_vlan_end=3999
  cmms_per_rack=8
  ```

- Enter the `configure-cluster` command, and use the menu-driven tool to add this information.

  The following is the path through the GUI:

  **I Initial Setup Menu > N Network Settings > M Configure Management Network VLAN Setting**

- Run the `cadmin` command after the `configure-cluster` command has been run but before `cmcdetectd` is enabled.

  The following are example `cadmin` commands:

  ```
  cadmin --set-mgmt-net-routing-protocol ospf
  cadmin --set-rack-start-number 1000
  cadmin --set-mgmt-vlan-start 2000
  cadmin --set-mgmt-ctrl-vlan-start 3000
  ```

The following are some important HPE Cray EX attributes:

| Attribute | Significance |
|---|---|
| `mgmt_net_routing_protocol` | Dictates the routing protocol used when `cmcdetectd` configures management switches to share routing details between RIP and OSPF. Note that HPE Aruba branded switches only support OSPF. |
| `rack_start_number` | A value that must match the lowest cabinet number in your system. For example, if your HPE Cray EX system lowest cabinet is `x1000`, set this value to `1000`. |
| `mgmt_vlan_start` | Specifies where `cmcdetectd` starts as the untagged VLAN. Using the suggested assignment, a `rack_start_number` of `1000` dictates a `mgmt_vlan_start` value of `2000`. |

*Table Continued*

| Attribute | Significance |
|---|---|
| mgmt_vlan_end | Specifies the end value of mgmt_vlan_start. The cmcdetectd daemon does not assign VLANs above this value, and this value must be higher than mgmt_vlan_start. |
| mgmt_ctrl_vlan_start | Specifies where cmcdetectd starts as the tagged VLAN. Using the suggested assignment from above, a rack_start_number of 1000 dictates a mgmt_ctrl_vlan_start value of 3000. |
| mgmt_ctrl_vlan_end | Specifies the end value of mgmt_ctrl_vlan_start. The cmcdetectd daemon does not assign VLANs above this value, and this value must be higher than mgmt_ctrl_vlan_start. |
| cmms_per_rack | Specifies the number of HPE Cray EX chassis management modules (CMMs) that exist in a cabinet to the cluster manager. This value is unlikely to change from the default of 8. |

## Chassis controllers and VLANs on an HPE Apollo 9000 cluster

The HPE Apollo 9000 system chassis controller default values are as follows:

- The default untagged VLAN numbering begins at 2001. The end number depends on how many chassis controllers are allowed in a management VLAN. By default, 8 chassis controllers are allowed per management VLAN. By default, the VLAN begins at VLAN 2001 and ends at VLAN 2999.
- The HPE Apollo 9000 chassis controllers do not used a tagged cooling VLAN.

The following commands retrieve VLAN information or configure VLAN settings:

- To view how many chassis controllers are allowed in a management VLAN, use the following command:

  ```
  # cadmin --show-cmcs-per-mgmt-vlan
  8
  ```

- The number of chassis controllers allowed in a management VLAN can be one of the following:
  - 0.

    Or
  - An integer that is a multiple of 4 and no greater than 48.

  To change how many chassis controllers 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.

  **NOTE:** On HPE Apollo 9000 systems, when *number* is set to 0, it disables the automatic generation of VLANs for chassis controllers.

  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
  ```

## Reviewing the chassis controller configuration

The following procedure explains how to review the current chassis controller 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) Use the `switchconfig unset` command to reset the management switch ports.

   If you know the management switch and the ports to which a chassis controller 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 chassis controller.

   Example: Assume that rack 1, chassis controller 1 is plugged into `mgmtsw0` ports 1/0/11 and 2/0/11. Issue the following command to set both ports back to default settings:

   ```
   # switchconfig unset --switches mgmtsw0 --ports 1/0/11 --redundant yes
   ```

5. Flip the power breakers on for the chassis controllers, one rack at a time.

   Notice that `cmcdetectd` runs in a serial fashion. It handles one chassis controller configuration at a time to prevent configuration conflicts on the management switches.

6. Verify the configuration.

   After the `cmcdetectd` service configures a chassis controller, the system logs an entry to the following file:

   `/etc/cmc-switch-info.txt`

Verify that this file contains the correct entries.

Example 1. The chassis controller entry might look as follows:

```
# cat /etc/cmc-switch-info.txt
mac_address=00:fd:45:ff:3b:46, mgmtsw=mgmtsw0, vlans=None, default_vlan=2001, bonding=manual,
ports=1/0/11, redundant=yes, cmc_type=nonice, cmc_hostname=r1c1
# VLAN to management switch configuration
vlan=2001, mgmtsw=mgmtsw0, configured=no, chassis_type=nonice, vlan_type=multinet
```

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 controller switches manually

To reapply the chassis controller 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.

---

**NOTE:** To configure L3 VLAN routing settings, you might need to change the `configured` VLAN property to `no` in order for `cmcdetectd` to re-apply a configuration to a management switch if the management switch has been reset. Example:

`vlan=####, … configured=no, …`

---

### Procedure

1. Make sure that all chassis controllers are configured.
2. Run the following command to configure the chassis controller switches:

```
# cmcdetectd --switchconfig
=== Reading the CMC-Switch configuration file: /etc/cmc-switch-info.txt ===

=== Running switchconfig for VLANs ===

VLAN 101 on management switch mgmtsw0 is an ICE VLAN, skipping L3 configuration

=== Running switchconfig for all CMCs ===

configuring CMC 08-00-69-16-C0-12 on managemenent switch mgmtsw0...
command: switchconfig set --switches mgmtsw0 --ports 1/11 --default-vlan 101 --bonding manual --redundant yes --vlans 3

saving configuration on management switch(es) mgmtsw0...
command: switchconfig config --switches mgmtsw0 --save

=== Results ===

Component               Function            Result
------------------------------------------------------------
CMC 08-00-69-16-C0-12    switchconfig set    successful
```

## Method 2 - Configuring the chassis controller 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 chassis controllers reside.
- The physical ports and management switches to which the chassis controllers are cabled.

If you cannot provide this information, do not use this method to configure the chassis controllers.

### 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 the `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 chassis controller.

   For example, assume that chassis controller `r1i0c` (that is, rack 1 chassis controller 0 using VLAN 101) is connected to management switch `mgmtsw0` on port `1:11` and `mgmtsw0` port `2:11`. 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 set` command, enter the following:

   ```
   # switchconfig set --help
   ```
4. Flip the power breakers on for the chassis controllers, 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 chassis controllers shortly after the chassis controllers are powered on.
5. (Optional) Validate the chassis controller configuration.

   Enter the following `switchconfig` command to display the bonding and VLAN configuration of the chassis controllers 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 1:11 with the following ports in a port-channel: 1:11, 2:11 in bonding mode: address-based L2
   ```

# Node provisioning takes too long or fails to complete

### Symptom

Node provisioning (or imaging) takes too long or fails to complete.

### Cause

If you use UDPcast, and provisioning takes too long or fails to complete, consider using BitTorrent or `rsync`. Sometimes a different file transfer method helps a node discovery command to complete more quickly.

### Action

1. Review the information in this step and select a different file transport method.

When you 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:

- A `cm node add` or a `cm node discover add` command completes and returns you to the system prompt.
- The leader nodes and 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`.

Regardless of the cluster type or node type, the default transfer method is BitTorrent. Other file transfer methods are `rsync` and UDPcast.

**Table 5: Compute node characteristics and image information** shows the node types and includes information about file transfer methods that are appropriate for each node.

### Table 5: Compute node characteristics and image information

| | Compute nodes with root disks | Compute nodes with NFS root file systems | Nodes with `tmpfs` root file systems |
|---|---|---|---|
| Transport path | The admin node installs the flat compute nodes.<br><br>Uses UDPcast, BitTorrent, or `rsync`. | On clusters with leader nodes, each node uses NFS to mount a root file system from its leader node. | 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. |
| Boot persistent? | Yes. | For compute nodes, boot persistence is possible depending on the configuration. | No.<br><br>All is lost on reboot. |
| Node image memory use | No. | For compute nodes, node image memory use depends on the configuration. | Yes.<br><br>The root file system consumes system memory. |

*Table Continued*

| | Compute nodes with root disks | Compute nodes with NFS root file systems | Nodes with `tmpfs` root file systems |
|---|---|---|---|
| RPM installation notes | N/A | For 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 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 6: File transfer methods** shows the available file transport methods.

### Table 6: File transfer methods

| | `rsync` | BitTorrent | UDPcast |
|---|---|---|---|
| Status | Not default | Default | Not 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.<br><br>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 | Yes | Yes |

*Table Continued*

| | rsync | BitTorrent | UDPcast |
|---|---|---|---|
| Appropriateness | Suited for a small number of nodes (2-5). If you have many nodes, run a node discovery command multiple times, and target different groups of nodes each time. | Suited for a large number of nodes. | Most efficient for large numbers of nodes.<br><br>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`.<br><br>The system uses `rsync` to transfer the miniroot file system | The miniroot file system. |

In addition to the tables, you can use the following figure to help you select a transport method:

**Figure 5: Selecting a transport method for provisioning clusters that have leader nodes**

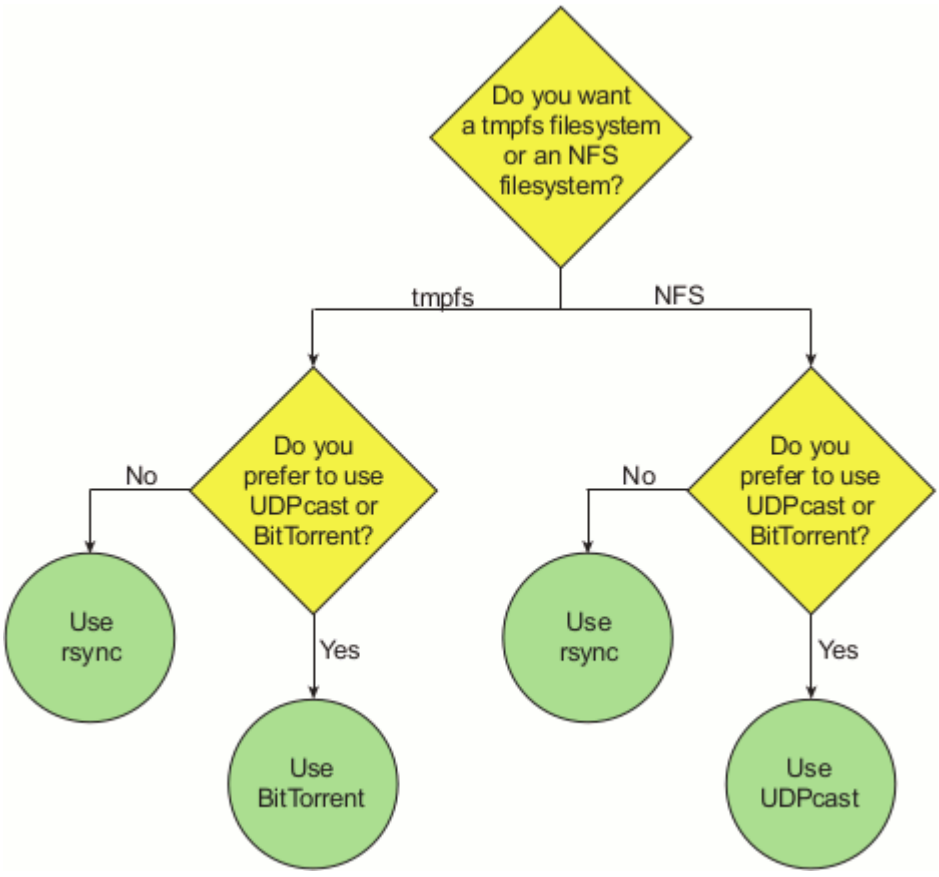


**Figure 5: Selecting a transport method for provisioning clusters that have leader nodes**

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`

3. Run the `cm node add` 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 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 [--transport` *`method`*`] -n` *`failed_node_name`*

   For *method*, specify one of the following data transport methods:

   - `bittorrent`
   - `rsync`
   - `udpcast`

   If you continue to have problems, you might have to change the transport method again. If UDPcast and BitTorrent both fail, specify `rsync`.

   Your site network configuration can affect the speed at which the `cm node provision` 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 2021-06-26 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:

   ```
   2021-05-03 14:16:07+0000 [-] Log opened.
   2021-05-03 14:16:07+0000 [-] twistd 14.0.2 (/usr/bin/python 2.7.9) starting up.
   2021-05-03 14:16:07+0000 [-] reactor class: twisted.internet.pollreactor.PollReactor.
   2021-05-03 14:16:07+0000 [-] Changing process name to clmgr-power
   2021-05-03 14:16:07+0000 [-] Log opened.
   2021-05-03 14:16:07+0000 [-] twistd 14.0.2 ( 2.7.9) starting up.
   2021-05-03 14:16:07+0000 [-] reactor class: twisted.internet.pollreactor.PollReactor.
   2021-05-03 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.

# 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
```

# Nodes are taking too long to boot

**Symptom**

Nodes are taking too long to boot.

**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.

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 `cm node add` command or the `cm node discover add` command has run.
- The node in question is configured into the cluster.
- You need to change the bonding mode for the node.

The following procedure works on any type of node, including an admin 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:

   `cm node show --mgmt-bonding -n` *hostname*

   For *hostname*, specify the hostname of the node you want to verify.

   Example:

   ```
   admin~# cm node show --mgmt-bonding -n leader1
   active-backup
   ```

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:

   ```
   admin# cm node set -n n0 --mgmt-bonding 802.3ad
   ```

   Example:

   ```
   admin~# cm node set -n leader1 --mgmt-bonding 802.3ad
   ```

4. Reset the 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:

- `node` for compute nodes.

  Or

- `leader` for scalable unit (SU) leader nodes.

For *hostname*, specify a node hostname.

Example:

```
admin~# cm power reset -t node n0
```

Example:

```
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:

   ```
   # switchconfig_configure_node --node leader1
   ```

# Nodes fail to boot

**Symptom**

A node can fail to boot for many reasons. This topic explains one remedy, which is to try booting the node with GRUB 2, rather than the default of iPXE.

For additional information, see the following:

**`dhcp_bootfile`**

**Action**

1. Log into the admin node as the root user.
2. Use the following command to verify whether the node is enabled to load iPXE:

   ```
   cm node show --dhcp-bootfile -n node
   ```

   For *node*, specify the hostname of the node that did not boot.
3. Use the following command to specify that iPXE load first and that iPXE load GRUB version 2:

   ```
   cm node set --dhcp-bootfile method -n node
   ```

   Select one of the following for *method*:

| *method* | Recommendation |
|---|---|
| `grub2` | Loads `grub2` directly. Specify `grub2` in most situations.<br><br>The `grub2` specification is the only specification supported on Arm (AArch64) platforms. |
| `ipxe-direct` | Specify if `grub2` and `ipxe` both fail to boot the node. This *method* uses a special iPXE binary on UEFI and legacy BIOS systems to directly load the kernel and `initrd`. It avoids `grub2`.<br><br>For example, HPE Cray EX systems require `ipxe-direct` for compute nodes.<br><br>Default. |

4. Boot the node.

   If the node still does not come up, the following are some additional actions you can take:

   - Attach a console or review the console log.
   - If node is in a shell, check its date.
   - Inspect the node entry in the cluster definition file for missing or incorrect parameters.
   - Verify that BIOS settings are correct.

# Cannot find the management switch that a node is plugged into

### Symptom

You 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.

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.

# Log files

Some log files reside in the `/var/log` directory and the `/opt/clmgr/log` directory.

On the admin node, the `/var/log/messages` file is one of the important log files.

On cluster with leader nodes, the `/var/log/dhcpd` file is another important log file.

The following are some other log files in the `/var/log` directory that might interest you:

- `/var/log/cmcdetectd.log`

  On the admin node, `cmcdetectd` logs its actions as it configures the switches for chassis controllers in the system. Watch for progress or errors here.

- `/var/log/dhcpd`

  This file contains DHCP messages.

  This file resides on the admin node.

  On clusters with leader nodes, a file by this name also resides on each leader node.

- `/var/log/switchconfig.log`

  On the admin node, there is a `switchconfig` command-line tool. This tool is largely used as nodes are configured into the cluster. 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/*`

# 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.

  In the 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.
- To set the hardware clock to the system clock, enter the following command:

  ```
  # hwclock --systohc
  ```

- To set the hardware clock to the system clock on the leader nodes, enter the following command:

  ```
  # clush -g leader hwclock --systohc
  ```

- To set the hardware clock to the system clock on the compute nodes, enter the following command:

  ```
  # clush -g compute hwclock --systohc
  ```

- To confirm the current hardware clock time, enter the `hwclock` command without options, as follows:

  ```
  # hwclock
  Thu 26 Jan 20XX 10:57:27 PM CST  -0.750431 seconds
  ```

- To confirm the current hardware clock on leader nodes, enter the following command:

  ```
  # clush -b -g leader date
  r1lead: Tue Apr 18 15:00:20 PDT 20XX
  ```

- To confirm the current hardware clock on compute nodes, enter the following command:

  ```
  # clush -b -g compute date
  node0: Tue Apr 18 15:00:45 PDT 20XX
  ```

# Switch wiring rules

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 configuring cascaded switches, it is impossible to know the connected switch ports of all trunks in advance, so 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
  - Compute nodes with services installed upon them
  - Chassis controllers

  Examples:

- If you connect the first NIC on leader node `leader1` to switch A, port 43, connect the second NIC on leader node `leader1` to switch B, port 43.
- If you connect chassis controller `r1i0c` chassis controller 0 port to switch A, port 2, then `r1i0c` chassis controller 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 configuring cascaded switch 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 configured into the cluster database, leave one trunk leg unplugged temporarily to prevent looping.
  - The node discovery commands tell 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.

# Miniroot operations

The following topics can help you troubleshoot a suspected miniroot kernel problem:

- **Miniroot functioning**
- **Entering rescue mode**
- **Logging into the miniroot to troubleshoot an installation**

## Miniroot functioning

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
- Compute nodes

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 `n0`. The following command logs you into the miniroot on node `n0` from the admin node:

```
admin# ssh -p 40 root@n0
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.

## 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

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.

   Log into the admin node if your goal is to restart UDPcast services for compute nodes.

2. Use one of the following commands to stop the `systemimager-server-flamethrowerd` services:

   From an admin node or a highly available (HA) admin node, enter the following command:

   ```
   # systemctl stop systemimager-server-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.

Use one of the following commands:

From an admin node or an HA admin node, enter the following command:

```
# systemctl start systemimager-server-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 mgmtsw`*`X`* `--info`

- To the IGMP status for a specific VLAN on a management switch:

  `switchconfig igmp --switches mgmtsw`*`X`* `--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 mgmtsw`*`X`* `--disable`

- To disable IGMP on a specific VLAN on a management switch:

  `switchconfig igmp --switches mgmtsw`*`X`* `--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
```

# Troubleshooting the `cmcinventory` service on an HPE Cray EX cluster

### Symptom

Use the troubleshooting information in this topic any time the cluster fails to return information about a node status or fails a power command. For example, if you enter the `cm node show` command, and some nodes are missing from the output, you can use the information in this topic to troubleshoot the `cmcinventory` service.

### Action

1. Enter the following command to determine whether the missing nodes are still in the cluster database:

   ```
   # cm node show -n "*" --not-exist
   ```

   If the nodes were physically removed from the cluster at one time, the cluster database might still include information for the nodes. This command returns a list of node hostnames that are not physically attached to the cluster but have been retained in the cluster database.

   If this command returns nothing, then the nodes are not in the cluster database. In this case, use the `cm node add` command to add the node to the cluster.

2. Verify the node entry in the `fastdiscover.conf` file:

   ```
   # grep x1006c6s2b1 /opt/clmgr/cmcinventory/conf/fastdiscover_mt.conf
   internal_name=x1006c6s2b1,hostname1=x1006c6s2b1,mgmt_bmc_net_macs="02:03:ee:06:32:10",
   mgmt_bmc_net_name=hostctrl3006,rack_nr=1006,chassis=6,tray=2,cmm_parent=x1006c6,...
   ```

3. Verify that entries in the cluster database match your knowledge of the node:

   ```
   # cm controller show -c x1006c6s2b1
   NAME        IPADDRESS     MACADDRESS        PROTOCOL CHANNEL USERNAME PASSWORD
   x1006c6s2b1 10.176.24.138 02:03:ee:06:32:10 None     None    root     U2FsdGVkX1+CYezJpHqf7qAigMwS4Vjsr28b+VpjkS0=
   # cm controller show -l -c x1006c6s2b1
   NAME         RACK  CHASSIS  TRAY  NODE
   x1006c6s2b1  None  6        2     1006
   ```

4. Examine the errors in the `/opt/clmgr/cmcinventory.log` file.

   For example, assume that the following entry exists in the log file:

   ```
   2021-09-23T10:35:43-0500 [Uninitialized] 'ERROR: x1006c6s2b1: trest_nc_obj_status - \
   An error occurred while connecting: 113: No route to host.'
   ```

   a. Verify whether the node is reachable:

      ```
      # ping x1006c6s2b1
      ```

   b. (Conditional) Verify the node status and power.

      If you are working with an HPE representative, you can use the `rest_agent_tool` command.

      ```
      # rest_agent_tool -db x1006c6 --controller_status -U root -P initial0 | grep Blade2
      PATH: https://x1006c6/redfish/v1/Chassis/Blade2
      ('Power Status:', 'x1006c6', 'Chassis/Blade2', u'On')
      ```

      **NOTE:** Use this command only while working with an HPE representative. for more information, enter the following:

      ```
      # rest_agent_tool -h
      ```

# Troubleshooting the `cmcinventory` service on an HPE Apollo 9000 cluster

You can examine the contents of the inventory files that the cluster manager uses when it automatically configures 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 controller IP address or the chassis controller 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. Use the `crm_mon` command to determine which physical node hosts the virtual admin node at this time.
3. Log into the node that hosts the virtual admin node, and enter the following command in a terminal window:

   ```
   # virsh console sac
   ```

# Nodes configured but with mismatched BIOS settings

**Symptom**

This problem presents itself when nodes that are identical, or are presumed to be identical, exhibit different behaviors. For example, some nodes might PXE boot and others might not.

You can use the remedy in this topic to analyze BIOS differences in the following additional circumstances:

- To compare the BIOS differences between two or more nodes. For example, you can see the boot order for different nodes easily.
- To retrieve information about node differences related to Hyperthreading or other settings.

In addition to using the commands in this topic for troubleshooting, you can use these commands to adjust BIOS settings for performance.

**Cause**

New nodes were configured into the cluster, but the BIOS settings on the new nodes do not match the BIOS settings on the other nodes.

**Action**

1. Log into the admin node as the root user.
2. Use the `cm node show` command to display the cluster nodes.

   For example:

   ```
   # cm node show
   n1
   n2
   n3
   n4
   ```

3. Use the `cm node bios show` to display the BIOS setting differences.

   The format is as follows:

   ```
   cm node bios show -n nodes --cmdiff
   ```

   For *nodes*, specify the node hostnames.

   You can use a mouse to click on the highlighted lines in the output to display the differences for each node.

   For example:

   ```
   # cm node bios show -n n2,n[3-4] --cmdiff
   ```

4. Adjust the BIOS settings as needed.

   The cluster manager provides the following commands:

   - `cm node bios show`, which shows BIOS settings for nodes.
   - `cm node bios set`, which lets you set BIOS features.
   - `cm node bios reset`, which lets you reset the BIOS to factory settings.

   Use the BIOS documentation and HPCM to adjust the BIOS settings.

# Cluster manager cannot find a suitable disk

**Symptom**

You began to install the cluster manager and selected a boot option from the **Display Instructions** menu. The miniroot exited and displayed instructions that describe how to proceed. These actions occur when the cluster manager cannot determine the disk to use for the installation. The possibilities are as follows:

- If you reinstall the cluster manager, the cluster manager overwrites existing disks with new cluster manager data and labels. In this case, the cluster manager recognizes the existing disks as belonging to a previous cluster manager installation. The installation proceeds as expected without issuing any messages.
- If you reinstall the cluster manager onto disks that were not part of a previous cluster manager installation, the cluster manager does not recognize the disks. In this case, the cluster manager miniroot exits and issues instructions in a message.
- If you start an installation, or a reinstallation, and there is more than one blank disk device, the cluster manager does not arbitrarily choose a disk. In this case, the cluster manager miniroot exits and issues instructions in a message.

If your console terminal did not buffer the instructional messages, you can find the messages in the following file:

`/tmp/si.log`

Before proceeding to the solutions that follow, complete the following prerequisites:

1. Copy or write down all disk devices in the following directory:

   `/dev/disk/by-path/`

2. Select a disk to use for the installation.

---

**NOTE:** None of the solutions that follow enable you to configure the target disk as a BIOS-assisted software RAID. The cluster manager does not support a target disk with that configuration.

## Solution 1

### Cause

No blank disk was found.

### Action

1. Enter the following command:

   `sgdisk --zap-all /dev/disk/by-path/`*`target_device_name`*

2. Reset the system.
3. Start the installation again.

## Solution 2

### Cause

Too many blank disks were found, and you expected the disks to be configured into a hardware RAID.

### Action

1. Reset the system.
2. When prompted, enter the RAID controller.
3. Use documentation for the RAID to configure the RAID controller.
4. Start the installation again.

### Solution 3

#### Cause

Too many blank disks were found, but you know the exact disk you want the cluster manager to use for the installation.

#### Action

Reset the system, and add the following command to the parameter passed to the installer:

```
force_disk="/dev/disk/by-path/target_device_name"
```

### Solution 4

#### Cause

Too many blank disks were found, and you want an MD RAID array.

#### Action

Reset the system, and add the following RAID10 commands to the parameter passed to the installer:

```
md_metadata=md, \
md_raidlevel=10, \
force_disk="/dev/disk/by-path/1st_target_device_name, \
/dev/disk/by-path/2nd_target_device_name, \
/dev/disk/by-path/3rd_target_device_name, \
/dev/disk/by-path/4th_target_device_name"
```

## Socket failure when connecting to the configuration manager

#### Symptom

A socket failure can occur when you attempt the following:

- You run the following command:

  ```
  cm node update config --sync -n admin
  ```

- You change the IP address of the admin node.

In the preceding cases, the cluster manager might issue the following messages:

```
ERROR: Socket failure connecting to configuration manager ('172.xx.xx.xx', 1030): Connection refused
ERROR: Retrying in 0.500 seconds
ERROR: Socket failure connecting to configuration manager ('172.xx.xx.xx', 1030): Connection refused
ERROR: Failed to contact configuration manager
```

#### Cause

These messages reflect a failure to connect with the configuration manager.

**Action**

1. Log into the admin node as the root user.
2. Enter the following command to restart the configuration manager service:

   ```
   # systemctl restart config_manager.service
   ```

# Replacing and servicing nodes

You can install and configure a spare node to replace failed system disks.

The failed node can be any kind of node, including an admin 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.

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 node or a compute node.

**NOTE:** If you are using multiple root slots, the installation procedures affect only the current slot.

For information about other hardware operations and about replacing other types of cluster components, see the following:

**HPE Performance Cluster Administration Guide**

## Replacing HPE Cray EX compute nodes

The procedure in this topic explains how to remove compute nodes from an HPE Cray EX cluster and replace them with new compute nodes. This procedure assumes that you want to remove all information about the old nodes from the cluster database. As a result, the new nodes can appear in the cluster database with new IP addresses and new MAC addresses.

**Procedure**

1. Log into the admin node as the root user.
2. Enter the following command to find out if the `cmcinventory` service is running:

   ```
   # systemctl status cmcinventory.service
   ```

3. Open file `/opt/clmgr/etc/cmcinventory.conf`, and search for the `mac_update` field.

   Make sure that the field is set as `mac_update=True`, and then save and close the file.

4. (Conditional) Start the `cmcinventory` service.

   Complete this step if the `cmcinventory` service is not running.

   The command is as follows:

   ```
   # systemctl start cmcinventory.service
   ```

5. Power off the compute node or nodes you want to remove.

   Example 1. The following command powers off the compute nodes on blade `x1002c1s0b0`:

   ```
   # cm power off -t node "x1002c1s0b0*"
   ```

   Example 2. The following command powers off slot `x1002c1s0`:

   ```
   # cm power off -t slot x1002c1s0
   ```

6. Complete the physical tasks needed for these nodes.
7. Power on the compute node or nodes you worked on.

Example 1:

```
# cm power on -t slot x1002c1s0
```

Example 2:

```
# cm power on -t node "x1002c1s0b0*"
```

8. Observe the messages in the `cmcinventory` log file:

   ```
   # tail -f /opt/clmgr/log/cmcinventory.log
   ```

   The `cmcinventory` service powers up all components, configures controllers, and then configures the nodes.

9. Display the status of the nodes:

   Example 1:

   ```
   # cm power status -t node "x1002c1s0b0*"
   ```

   Example 2:

   ```
   # cm power status -t slot x1002c1s0
   ```

   Upon success, this command indicates that all nodes are booted.

   If the power status does not show the nodes as booted, see the following:

   **Troubleshooting the `cmcinventory` service on an HPE Cray EX cluster**

10. Continue monitoring the `cmcinventory` log file, and stop the `cmcinventory` service when you start to see messages with `Changed 0` at the end, such as the following:

    ```
    1-07T13:13:51+0000 [stdout#info] 2021-01-07 13:13:51,089 - admin.cmminv - DEBUG - Done Node Controller Status
    2021-01-07T13:13:51+0000 [stdout#info] 2021-01-07 13:13:51,089 - admin.cmminv - DEBUG - Load Node Mac Addrs
    2021-01-07T13:13:52+0000 [stdout#info] 2021-01-07 13:13:52,224 - admin.cmminv - DEBUG - Done Node Mac Addrs
    2021-01-07T13:13:52+0000 [stdout#info] 2021-01-07 13:13:52,225 - admin.cmminv - DEBUG - Populate SCI inventory jsons
    2021-01-07T13:13:52+0000 [stdout#info] 2021-01-07 13:13:52,225 - admin.cmminv - DEBUG - Compare Inventory CMMs 2,
    Changed 0


    # systemctl stop cmcinventory.service
    ```

11. (Conditional) Restart the `cmcinventory` service.

    Complete this step if you want a fresh hardware scan or if a hardware change was not detected.

    Enter the following commands:

    a. Stop the `cmcinventory` service:

       ```
       # systemctl stop cmcinventory
       ```

    b. Remove the existing inventory file and initiate a new scan:

       ```
       # rm /opt/clmgr/cmcinventory/inventory/inventory.xhardware
       ```

       For *hardware*, specify the identifier for the hardware you replaced.

    c. Start the `cmcinventory` service:

       ```
       # systemctl start cmcinventory
       ```

    d. Monitor the log file for updates:

       ```
       # tail -F /opt/clmgr/log/cmcinventory.log
       ```

# Servicing HPE Cray EX compute nodes

The procedure in this topic explains how to remove a compute node from an HPE Cray EX cluster in a way that preserves the node IP address and the node MAC address. For example, you can use this procedure to replace a power supply in a node.

**Procedure**

1. Log into the admin node as the root user.
2. Open file `/opt/clmgr/etc/cmcinventory.conf`, and search for the `mac_update` field.

   Make sure that the field is set as `mac_update=True`, and then save and close the file. This setting ensures that the node IP address and MAC address information is retained in the cluster database when you physically remove the node from the cluster.
3. Power off the slot in which the node resides.

   For example, assume that you want to service node `x1002c1s0b0n0`. Power off slot `s0`:

   ```
   # cm power off -t slot x1002c1s0
   ```
4. Complete the physical tasks needed for the node.
5. Power on the slot in which the node resides.

   For example:

   ```
   # cm power on -t slot x1002c1s0
   ```
6. Observe the messages in the `cmcinventory` log file:

   ```
   # tail -f /opt/clmgr/log/cmcinventory.log
   ```

   The `cmcinventory` service powers up all components, configures controllers, and then configures the nodes.
7. Display the status of the slot:

   ```
   # cm power status -t slot x1002c1s0
   ```

   Upon success, this command indicates that all nodes in the slot are booted.

   If the power status does not show all the nodes as booted, see the following:

   **Troubleshooting the `cmcinventory` service on an HPE Cray EX cluster**

# Replacing a node

The procedure in this topic explains how to replace an entire node, including the system disks in the node. This procedure does not preserve the original system disks.

For example, you can use this procedure to replace service nodes on any cluster.

You can use the procedure in this topic to replace nodes that are not automatically managed by any of the following:

- `cmcdetectd`
- `cmcinventory`

Do not use this procedure for the following:

- HPE Cray EX compute nodes
- HPE Apollo 9000 compute nodes

**Procedure**

1. Verify that you have an appropriate spare node.

   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.

   The following are some reasons to have the two types of spares:

   - Admin node BIOS settings are different from BIOS settings for other nodes.

     For example, the boot order is different for each node type.

     Attempts to configure the node into a cluster will fail.
   - Depending on your site policy, the node controllers of an admin node might or might not be configured to use DHCP by default.
   - The management cards of leader nodes and compute nodes must be configured to use DHCP by default.

     Otherwise, attempts to configure the node into the cluster will fail.

2. Connect a keyboard, video screen, and mouse to the node.
3. If possible, power down the failed node.
4. 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.
5. Disconnect all power cables.
6. 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.
7. Remove any peripheral components, such as a keyboard, video screen, or mouse, from the node.
8. Remove the failed node from the rack.
9. Install the shelf spare node into the rack.
10. Connect the Ethernet cables in the same way they were connected to the replaced node.
11. Connect AC power.
12. Connect to the node through the node controller or attach a keyboard, video screen, and mouse to the node.
13. From the admin node, update the following in the cluster manager database:

- The MAC address of the spare
- The MAC address of the node controller 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 node controller 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:

The following example displays the MAC address of 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:

The following example sets the MAC address of `n0`:

```
# cm node set --mac-address 00:25:90:04:4e:01 -n n0
```

Example:

The following example shows the MAC address of the node controller 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:

The following example sets the MAC address of the node controller on `n0`:

```
# cm node set --bmc-mac-address 00:25:90:03:51:1d -n n0
```

14. Power up the replaced node.

# Replacing failed system disks in a node that uses a disk drive for its root file system

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.

**NOTE:** Do not use this procedure to replace the Gluster disks in a scalable unit (SU) leader node.

### Procedure

1. Verify that you have an appropriate spare system disk.

   If necessary, obtain new system disks from HPE.

A spare system disk is equivalent to one of the system disks on your running cluster. The spare sits on a shelf. 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.

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 node controller 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:

| Variable | Specification |
|---|---|
| *hostname* | The hostname of the node with the new system disks. |
| *image* | 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. Verify that you have an appropriate spare node.

   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 types of spare nodes:

- One spare for the admin node.
- One spare for a compute node.

The following are some reasons to have the two types of spares:

- Admin node BIOS settings are different from BIOS settings for other nodes.

  For example, the boot order is different for each node type.

  Attempts to configure the node into a cluster will fail.
- Depending on your site policy, the node controllers of an admin node might or might not be configured to use DHCP by default.
- The management cards of leader nodes and compute nodes must be configured to use DHCP by default.

  Otherwise, attempts to configure the node into the cluster will fail.

2. Connect a keyboard, video screen, and mouse to the node.
3. If possible, power down the failed node.
4. 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.
5. Disconnect all power cables.
6. 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.
7. Remove any peripheral components, such as a keyboard, video screen, or mouse, from the node.
8. Remove the failed node from the rack.
9. Remove the system disks from the failed node.

   That is, open the failed node and remove the system disks.
10. Remove the system disks from the new node.

    That is, pull the current system disks, using their carriers, and set the disks aside.
11. Insert the preserved disks from the failed node into the new node (the shelf spare).
12. Insert the new node, with the preserved disks, back into the rack.
13. Connect AC power to the new node.
14. Connect a keyboard, video screen, and mouse to the new node.
15. 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 node controller 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 node controller 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 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 node controller 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 node controller on `n0`:

```
# cm node set --bmc-mac-address 00:25:90:03:51:1d -n n0
```

16. Power up the replaced node.
17. (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 a node**

   Despite the title, the steps in the preceding procedure apply to the task of replacing SU leader nodes.
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:

   **(Conditional) 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:

| Variable or parameter | Specification |
|---|---|
| `--reintegrate-whole-leader` | A required parameter. |
| `--destroy-gluster-disk` | (Optional) Use this parameter if there are partitions or information on the disk at this time. This parameter reformats the disk. |
| *hostname* | The hostname of the new SU leader node. |

6. Enter the following commands to monitor the Gluster rebalancing:

   - `ssh leader`*X*` gluster volume heal cm_shared info summary`

     The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.
   - `ssh leader`*X*` gluster volume heal cm_logs info summary`
   - `ssh leader`*X*` 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. Back up the cluster according to the procedure in the following topic:

   **Backing up the cluster**

2. For each new SU leader node, obtain the MAC address of each of the following:

   - The node controller MAC address
   - The Ethernet MAC address

   For information about how to retrieve the MAC address of the spare, see the documentation for the node controller for the spare.

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

4. Update the cluster definition file that contains information about 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 image in the `templates` section and about the three new SU leader nodes at the end:

```
# File suleader.config
# Cluster definition file for 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, bmc_username=ADMIN, bmc_password=ADMIN,
force_disk="/dev/disk/by-path/pci-0000:5c:00.0-scsi-0:1:0:0", image=su-rhel8.4,
kernel=4.18.0-305.el8.x86_64
[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=ib0, net_ifs="ib0"
template=su-leader, network=ib1, net_ifs="ib1"
[discover]
# 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.

5. Configure the new SU leader nodes into the cluster.

   For example, to configure the nodes described in `suleader.config`, enter the following command:

   ```
   # discover --update-templates --configfile suleader.config --all
   ```

6. Wait for the new nodes to come up.

   For example, if the node is up, you can `ssh` to the node.

7. 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:

   **(Conditional) Creating a scalable unit (SU) leader node list file**

8. 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_hostname3*, specify the hostnames of the three new SU leader nodes.

   For example:

   ```
   # su-leader-setup --add-leaders leader4,leader5,leader6
   ```

9. Complete one of the following procedures to confirm the SU leader setup:

   - **Configuring the scalable unit (SU) leader node software on an HPE Cray EX cluster or an HPE Apollo 9000 cluster**
   - **Configuring the scalable unit (SU) leader node software on an HPE Apollo cluster with SU leader nodes that is not an HPE Apollo 9000 cluster**

10. Enter the following commands to verify that there are Gluster volumes for all the SU leader nodes:

    - `ssh leader`*`X`*` gluster volume heal cm_shared info summary`

      The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.
    - `ssh leader`*`X`*` gluster volume heal cm_logs info summary`
    - `ssh leader`*`X`*` gluster volume heal ctdb info summary`
    - `ssh leader`*`X`*` gluster volume heal cm_obj_sharded 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**

11. Back up the cluster configuration again:

    **Backing up the cluster**

## Replacing a Gluster disk in a scalable unit (SU) leader node

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:

   **(Conditional) 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 leader`*X*` gluster volume heal cm_shared info summary`

     The `cm_shared` volume is the largest Gluster volume, and its healing time is longer.
   - `ssh leader`*X*` gluster volume heal cm_logs info summary`
   - `ssh leader`*X*` 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**

## Reinstalling the software on a scalable unit (SU) leader node

The procedure in this topic assumes that you have an installed cluster that includes SU leader nodes, but you need to reinstall the software on the SU leader nodes.

**Procedure**

1. Log into the admin node as the root user.
2. Enter the following command to decouple the admin node from the Gluster file system, and follow the instructions that the command returns:

   `# disable-su-leader --force`

   Make sure that the command completes successfully, as follows:

- When the command completes successfully, it returns a set of steps to ensure complete decoupling of the admin node from the SU leader. Follow all the steps shown in the output. For example:

  ```
  # disable-su-leader --force
  Unmounting bind-mounted content related to shared storage
  Unmounting shared storage itself
  umount: /opt/clmgr/shared_storage: not mounted.
  umount: /opt/clmgr/cm_logs: not mounted.
  umount: /opt/clmgr/cm_obj_sharded: not mounted.
  Removing shared storage bind entries and gluster client mount from /etc/fstab
  Steps you must take:
  .
  .
  .
  ```

- When the command does not complete successfully, it returns output similar to the following:

  ```
  # disable-su-leader --force
  Unmounting bind-mounted content related to shared storage stop and start conserver so that
  it can use local storage for consoles localhost umounts:
  umount /opt/clmgr/tftpboot/grub2/cm;
  umount /opt/clmgr/tftpboot/ipxe-direct/cm;
  umount /opt/clmgr/tftpboot/images;
  umount /opt/clmgr/image/images_ro_nfs;
  umount /opt/clmgr/image/images_rw_nfs;
  umount /opt/clmgr/image/torrents;
  umount /opt/clmgr/image/tarballs;
  umount /var/log/HOSTS;
  umount /var/log/consoles;
  umount /opt/clmgr/image/image_objects;
  umount /etc/dhcp/dhcpd.conf.d;
  umount: /var/log/consoles: target is busy.
  Unmounting shared storage itself
  Removing shared storage bind entries and gluster client mount from /etc/fstab 172.23.255.201:/cm_shared
  /opt/clmgr/shared_storage
  fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /opt/clmgr/image/images_rw_nfs fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /opt/clmgr/tftpboot/grub2/cm fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /opt/clmgr/image/images_ro_nfs fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /opt/clmgr/image/torrents fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /opt/clmgr/image/tarballs fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_logs
  /var/log/consoles fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0 172.23.255.201:/cm_shared
  /etc/dhcp/dhcpd.conf.d fuse.glusterfs
  rw,relatime,user_id=0,group_id=0,allow_other,max_read=131072 0 0

  Error: Shared storage mounts or bind mounts still found in /proc/mounts

  Please manually fix busy mount points and re-run this script again.
  !!! Action required !! System is in an inconsistent state now!!
  ```

  The preceding messages indicate that the command failed to unmount one or more locations. Unmount the locations manually and enter the `disable-su-leader --force` command again.

3. Use the `cm node provision` command in the following format to reinstall the software on the SU leader nodes:

   ```
   # cm node provision -l leader
   ```

4. Enter the following command, and verify that the expected leader nodes appear in the output:

   ```
   # cat /opt/clmgr/etc/su-leader-nodes.lst
   ```

5. Run the `su-leader-setup`, the `enable-su-leader`, and the `cm image activate` commands as shown in one of the following procedures:

- **Configuring the scalable unit (SU) leader node software on an HPE Cray EX cluster or an HPE Apollo 9000 cluster**
- **Configuring the scalable unit (SU) leader node software on an HPE Apollo cluster with SU leader nodes that is not an HPE Apollo 9000 cluster**

6. Reprovision the compute nodes:

```
# cm node provision -n 'node*'
```

# Upgrading from an HPE Performance Cluster Manager 1.x release

**Procedure**

1. **Starting the upgrade**
2. **Upgrading compute nodes**
3. **Upgrading the scalable unit (SU) leader nodes**
4. **Completing the upgrade**
5. **Additional upgrade procedures**

## 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. You do not need to complete this step to upgrade from HPE Performance Cluster Manager 1.3.X.

   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. You do not need to complete this step to upgrade from HPE Performance Cluster Manager 1.3.X.

   Enter the following command:

   ```
   # rm -rf /usr/share/logstash
   ```
6. Retrieve the `.iso` files you need from the Hewlett Packard Enterprise customer portal.

   Retrieve the upgrade `.iso` file for the HPE Performance Cluster Manager, and optionally, MPI and AIOps.

7. Complete the following steps to add new repositories:

   a. Add the cluster manager software:

      ```
      # crepo --add cm-1.6-*.iso
      ```

   b. Select the cluster manager repository.

      For example:

      ```
      # crepo --select Cluster-Manager-1.6-rhel84-x86_64
      ```

   c. Update the `crepo` packages.

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

      - On RHEL 8.X operating systems, enter the following command:

        ```
        # cinstallman --dnf-node --node admin update crepo crepo-libs
        ```

      - On RHEL 7.X operating systems, enter the following command:

        ```
        # cinstallman --yum-node --node admin update crepo crepo-libs
        ```

      - On SLES 15 SPX or SLES 12 SPX operating systems, enter the following command:

        ```
        # cinstallman --zypper-node --node admin update crepo crepo-libs
        ```

   d. 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-8.4-20200723.1-Server-x86_64-dvd1.iso
      ```

   e. (Optional) Add MPI software:

      ```
      # crepo --add hpe-mpi-1.9*.iso
      ```

   f. (Optional) Add the AIOps software.

      ```
      # crepo --add aiops-1.6-cd1-media-redhat-x86_64.iso
      ```

8. Create repository groups for any repositories you created.

   For example:

   ```
   # crepo --add-group hpcm-1.6 Red-Hat-Enterprise-Linux-8.4-x86_64 \
   Cluster-Manager-1.6-rhel84-x86_64 \
   HPE-MPI-1.9-rhel84-x86_64
   ```

9. (Conditional) Add the cluster manager AIOps repository to the HPE Performance Cluster Manager 1.6 repository group.

   Complete this step if you want to install AIOps.

   For example:

   ```
   # crepo --add-group hpcm-1.6 Cluster-Manager-AIOps-1.6-rhel83-x86_64
   ```

10. (Conditional) Remove old kernel versions to clear space in the `/boot` partition.

Complete this step if the admin node hosts a RHEL or CentOS operating system. Do not complete this step if the admin node hosts the SLES operating system.

It is possible that the `/boot` partition does not have enough free space for new kernel and `initramfs` images. This condition can cause the upgrade to fail and can cause new kernel RPMs to upgrade incorrectly.

Use one of the following methods to ensure that there is enough room in `/boot`:

Method 1 - Remove old kernels manually

**a.** Use the `df` command to determine the amount of free space left on the `/boot` partition.

For example:

```
# df -k /boot
Filesystem     1K-blocks   Used Available Use% Mounted on
/dev/sda11        289285  62576    207253  24% /boot
```

If the `Use%` column indicates 75% or more, you need to delete old kernels. Proceed to the next steps in Method 1.

**b.** Determine the current kernel version:

```
# uname -r
```

**c.** Display the list of installed kernels:

```
# rpm -q kernel
```

**d.** Remove specific old kernel versions.

RHEL 8.X example:

```
# cm node dnf -n admin remove kernel-4.18.0-240.el8.x86_64 \
kernel-core-4.18.0-240.el8.x86_64
```

RHEL 7.X example:

```
# cm node yum -n admin remove kernel-3.10.0-1160.el7.x86_64
```

Method 2 - Lower the `installonly_limit` setting to specify that fewer kernel versions are installed on future upgrades.

Open one of the following files, reset the value to be lower, and save and close the file:

- For RHEL 8.X admin nodes, edit file `/etc/dnf/dnf.conf`.
- For RHEL 7.X admin nodes, edit file `/etc/yum.conf`.

**11.** 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-1.6 \
  "dup --allow-vendor-change --auto-agree-with-licenses"
  ```

- 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-1.6 update yume
  # cinstallman --update-node --node admin --repo-group hpcm-1.6
  ```

- 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-1.6 update yume
  # cinstallman --update-node --node admin --repo-group hpcm-1.6
  ```

---

**NOTE:** At this point in the upgrade process, the HPE Performance Cluster Manager 1.6 command set is available. For continuity purposes, however, the upgrade instructions show the command set that was available in previous releases.

---

**12.** Refresh the admin node RPM list.

For example, the following command removes an RPM list in a repository group:

```
# crepo --recreate-rpmlists
```

**13.** Refresh the admin node.

For example:

```
# cinstallman --refresh-node --node admin --repo-group hpcm-1.6 \
--rpmlist /opt/clmgr/image/rpmlists/generated/\
generated-group-hpcm-1.6-admin.rpmlist
```

**14.** Use the `cm node refresh` command to refresh the secrets.

Refresh the secrets for the following nodes:

- Scalable unit (SU) leader nodes
- Compute nodes

This command requires that all nodes are online and accessible through `ssh`.

Example 1. To refresh the secrets for compute nodes `n0` through `n99`, enter the following command:

```
# cm node refresh secrets -n 'n[0-99]'
```

Example 2. To refresh node secrets for all nodes, enter the following command:

```
# cm node refresh secrets -n '*'
```

For more information about node input, enter `cm node refresh secrets -h`.

**15.** Run the following command:

```
# cm node update config -n '*' --sync
```

**16.** Run the following script:

```
# /opt/clmgr/lib/cluster-configuration
```

**17.** (Conditional) Restart the AIOps service.

Complete this step if you installed AIOps.

Enter the following command:

```
# cm aiops cooldev restart
```

18. Reboot the admin node.
19. Wait for the admin node to boot the operating system.
20. 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 1.6 upgrade"
    ```

    For *image*, specify the name of one of the node images.

# Upgrading compute nodes

**Procedure**

1. Upgrade the compute node image.

   This step differs depending on the operating system on the compute nodes.

   - For a SLES compute node image, enter a `cinstallman` command in the following format:

     ```
     cinstallman --zypper-image --image image --repo-group hpcm-1.6 \
     --duk "dup --allow-vendor-change"
     ```

   - For a RHEL 8.X or CentOS 8.X compute node image, enter commands in the following format:

     ```
     cinstallman --dnf-image --image image \
     --repo-group hpcm-1.6 update yume --duk
     ```

     and

     ```
     cinstallman --update-image --image image --repo-group hpcm-1.6 --duk
     ```

   - For a RHEL 7.X or CentOS 7.X compute node image, enter commands in the following format:

     ```
     cinstallman --yum-image --image image \
     --repo-group hpcm-1.6 update yume --duk
     ```

     and

     ```
     cinstallman --update-image --image image --repo-group hpcm-1.6 --duk
     ```

   For *image*, specify the name of one of the compute node images.

2. Refresh the compute node images.

   This step ensures that all the packages listed in the provided `rpmlist` file are actually installed. Enter the following commands:

   ```
   cinstallman --refresh-image --image image \
   --repo-group hpcm-1.6 \
   --rpmlist /opt/clmgr/image/rpmlists/generated/\
   generated-group-hpcm-1.6.rpmlist \
   ```

```
--duk

cinstallman --update-kernels --image image
```

3. Update the miniroot in the images to ensure that the miniroot is built from the latest available packages:

   ```
   cinstallman --update-miniroot --image image --recreate \
   --repo-group hpcm-1.6
   ```

4. (Conditional) Upgrade the compute nodes.

   Complete this step if the compute nodes are provisioned as `rootfs=disk` or `rootfs=tmpfs`. You do not need to complete this step if the compute nodes are provisioned to use an NFS root file system (`rootfs=nfs`).

   This step differs depending on the operating system of the compute nodes.

   - For SLES compute nodes, enter the following command:

     ```
     # cinstallman --zypper-node --node "n*" --repo-group hpcm-1.6 "dup --allow-vendor-change"
     ```

   - For RHEL 8.X or CentOS 8.X compute nodes, enter the following commands:

     ```
     # cinstallman --dnf-node --node "n*" --repo-group hpcm-1.6 update yume
     ```

     and

     ```
     # cinstallman --update-node --node "n*" --repo-group hpcm-1.6
     ```

   - For RHEL 7.X or CentOS 7.X compute nodes, enter the following commands:

     ```
     # cinstallman --yum-node --node "n*" --repo-group hpcm-1.6 update yume
     # cinstallman --update-node --node "n*" --repo-group hpcm-1.6
     ```

5. (Conditional) Refresh the compute node images.

   Complete this step if the compute nodes are provisioned as `rootfs=disk` or `rootfs=tmpfs`. You do not need to complete this step if the compute nodes are provisioned to use an NFS root file system (`rootfs=nfs`).

   For example:

   ```
   # cinstallman --refresh-node --node "n*" --repo-group hpcm-1.6 \
   --rpmlist /opt/clmgr/image/rpmlists/generated/\
   generated-group-hpcm-1.6.rpmlist
   ```

6. (Optional) Use the `cinstallman` command in the following format to update the kernels:

   ```
   cinstallman --assign-image --image image --kernel new_kernel \
   --node node ... node
   ```

7. (Conditional) Activate the NFS images.

   Complete this step for compute nodes that have NFS root file systems (`rootfs=nfs`).

   Enter the `cm image activate` command in the following format:

   ```
   cm image activate -i image [--delete-rw-content]
   ```

   For *image*, specify the image name.

   Include the `--delete-rw-content` parameter if the nodes use `nfs-overlay`.

8. Reset all nodes that received a new kernel or were otherwise upgraded with characteristics that require a reboot.

Enter the following command:

```
# cpower node reset "r*c*t*n*"
```

9. (Conditional) Set `CMCINVENTORY_MANAGED` for nodes managed by the `cmcinventory` service.

   Complete this step on HPE Cray EX clusters and on HPE Apollo 9000 clusters.

   Enter the following command:

   ```
   # for i in `cat /opt/clmgr/cmcinventory/conf/fastdiscover*.conf | awk -F 'hostname1=' '\
   { print $2 }' | cut -f1 -d ',' `; do /opt/sgi/sbin/cattr set cmcinventory_managed yes \
   --node $i; done
   ```

# Upgrading the scalable unit (SU) leader nodes

**NOTE:** Kafka and Elasticsearch can appear unhealthy during an upgrade. Some data might be unavailable. After the upgrade, the data becomes available again.

**Procedure**

1. Map your cluster.

   Make sure you know the SU leader Trios that work together, and obtain the hostnames of each SU leader node. If necessary, enter the following command to display the SU leader node hostnames:

   ```
   # cm node show -t role su-leader
   leader1
   leader2
   leader3
   leader4
   leader5
   leader6
   ```

   For example, assume that you have six SU leader nodes, grouped as follows, with hostnames in the format of `leader`*n*:

   - Trio 1:
     - `leader1`
     - `leader2`
     - `leader3`
   - Trio 2:
     - `leader4`
     - `leader5`
     - `leader6`

   Divide your leader nodes into Groups so that only one node of a Trio appears in one Group.

   For example, Group 1 can consist of the following SU leader nodes:

- leader1
- leader4

Group 2 can consist of the following SU leader nodes:

- leader2
- leader5

Group 3 can consist of the following SU leader nodes:

- leader3
- leader6

Your goal is to upgrade one Group at a time. In this way, only one SU leader node from one Trio is being upgrading at any time.

2. Upgrade the cluster trivial database (CTDB) packages.

   All leader nodes are required to run the same version of the CTDB software. If an operating system package update includes updating the CTDB software to a new version, the updated leader node might not be able to synchronize with the other leader nodes. The goal of this step is to update all the CTDB packages and minimize synchronization problems.

   An alternative, you can lock the CTDB package at a certain level. This action prevents a CTDB update at this time so you can update the CTDB software in the future at a time of your choosing. Explaining this alternative method is outside the scope of the cluster manager documentation.

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

   - On RHEL 8.X SU leader nodes, enter the following command:

     ```
     # cm node dnf -t role su-leader --repo-group hpcm16 update ctdb
     ```

   - On RHEL 7.X SU leader nodes, enter the following command:

     ```
     # cm node yum -t role su-leader --repo-group hpcm16 update ctdb
     ```

   - On SLES SU leader nodes, enter the following command:

     ```
     # cm node zypper -t role su-leader --repo-group hpcm16 update ctdb
     ```

3. Verify the CTDB status.

   For example:

   ```
   # ssh leader1 ctdb status
   pnn:0 172.23.0.3       OK (THIS NODE)
   pnn:1 172.23.0.2       OK
   pnn:2 172.23.0.4       OK
   pnn:3 172.23.0.5       OK
   pnn:4 172.23.0.6       OK
   pnn:5 172.23.0.7       OK
   Generation:1749430009
   Size:24
   ```

.
.
.

In the output, verify that the status for each leader node is **OK**. The preceding command output shows OK status for each leader.

4. Decide which SU leader node Group you want to upgrade at this time.

   For example, if you are starting the upgrade, select Group 1.

   Subsequent steps in this procedure require you to perform several upgrade steps on one Group, return to this step, and perform the upgrade steps on another Group. Eventually, all Groups are upgraded, and you can proceed to the last steps in this procedure.

5. Verify that Gluster file system is healthy.

   When the Gluster file system is healthy, the reboots needed during the upgrade process can complete without affecting the compute nodes attached to each SU leader node.

   For example, from the admin node, enter the following command to display information about Gluster health on all SU leader nodes:

```
# ssh leader1 gluster volume status
Status of volume: cm_logs
Gluster process                                 TCP Port RDMA Port Online Pid
------------------------------------------------------------------------------
Brick 172.23.100.11:/data/brick_cm_logs         49156    0         Y      70152
Brick 172.23.100.12:/data/brick_cm_logs         49154    0         Y      63021
Brick 172.23.100.13:/data/brick_cm_logs         49154    0         Y      63077
Self-heal Daemon on localhost                   N/A      N/A       Y      70356
Self-heal Daemon on 172.23.100.13               N/A      N/A       Y      62949
Self-heal Daemon on 172.23.100.12               N/A      N/A       Y      62890
Task Status of Volume cm_logs
------------------------------------------------------------------------------
There are no active volume tasks
Status of volume: cm_obj_sharded
Gluster process                                 TCP Port RDMA Port Online Pid
------------------------------------------------------------------------------
Brick 172.23.100.11:/data/brick_cm_obj_shar
ded                                             49157    0         Y      70164
Brick 172.23.100.12:/data/brick_cm_obj_shar
ded                                             49155    0         Y      63079
Brick 172.23.100.13:/data/brick_cm_obj_shar
ded                                             49155    0         Y      63136
Self-heal Daemon on localhost                   N/A      N/A       Y      70356
NFS Server on localhost                         2049     0         Y      70297
Self-heal Daemon on 172.23.100.13               N/A      N/A       Y      62949
NFS Server on 172.23.100.13                     2049     0         Y      63188
Self-heal Daemon on 172.23.100.12               N/A      N/A       Y      62890
NFS Server on 172.23.100.12                     2049     0         Y      63130
Task Status of Volume cm_obj_sharded
------------------------------------------------------------------------------
There are no active volume tasks
Status of volume: cm_shared
Gluster process                                 TCP Port RDMA Port Online Pid
------------------------------------------------------------------------------
Brick 172.23.100.11:/data/brick_cm_shared 49158    0         Y      70176
```

```
Brick 172.23.100.12:/data/brick_cm_shared 49152    0          Y       62827
Brick 172.23.100.13:/data/brick_cm_shared 49152    0          Y       62886
Self-heal Daemon on localhost             N/A      N/A        Y       70356
NFS Server on localhost                   2049     0          Y       70297
Self-heal Daemon on 172.23.100.13         N/A      N/A        Y       62949
NFS Server on 172.23.100.13               2049     0          Y       63188
Self-heal Daemon on 172.23.100.12         N/A      N/A        Y       62890
NFS Server on 172.23.100.12               2049     0          Y       63130
Task Status of Volume cm_shared
------------------------------------------------------------------------------
There are no active volume tasks
Status of volume: ctdb
Gluster process                           TCP Port RDMA Port Online Pid
------------------------------------------------------------------------------
Brick 172.23.100.11:/data/brick_ctdb      49159    0          Y       70183
Brick 172.23.100.12:/data/brick_ctdb      49153    0          Y       62963
Brick 172.23.100.13:/data/brick_ctdb      49153    0          Y       63020
Self-heal Daemon on localhost             N/A      N/A        Y       70356
Self-heal Daemon on 172.23.100.13         N/A      N/A        Y       62949
Self-heal Daemon on 172.23.100.12         N/A      N/A        Y       62890
Task Status of Volume ctdb
------------------------------------------------------------------------------
There are no active volume tasks

Expand (65 lines)
```

The preceding output shows a healthy Gluster file system. In your output, verify the following:

- All volumes should list all leader nodes.
- All leader nodes should have one `Brick`. In this example, the first three lines of this output show that.
- All bricks should have a `Pid` number.

**6.** Determine whether or not the upgrade includes updates to the Gluster file system software.

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

- On RHEL 8.X SU leader nodes, enter the following command:

  ```
  # cm node dnf -n leader1 check-update
  ```

- On RHEL 7.X SU leader nodes, enter the following command:

  ```
  # cm node yum -n leader1 check-update
  ```

- On SLES SU leader nodes, enter the following command:

  ```
  # cm node zypper -n leader1 update --dry-run --download-only
  ```

If the command output mentions `gluster` or `glusterfs`, then the upgrade includes Gluster updates.

**7.** (Conditional) Suspend CTDB operations during the upgrade.

Complete this step if Step **6** indicated that the upgrade includes Gluster updates.

The high availability (HA) monitoring scripts that are part of CTDB cannot be running while Gluster updates are running because of the following reasons:

- The scripts misinterpret the necessary restart of `glusterd`.
- The scripts take actions that conflict with the Gluster update process.

Complete the following steps:

**a.** Enter the following command to disable CTDB on the Group:

```
# pdsh -w leader1,leader4 systemctl stop ctdb
```

**b.** Confirm that only the leaders in the current Group are disconnected from CTDB. The goal of this step is to ensure that CTDB is not disabled for all the Groups at this time. You want only the Group undergoing upgrade at this time to have CTDB disabled. The format of the command to enter is as follows:

```
ssh leader_hostname ctdb status
```

For *leader_hostname*, enter the hostname of a leader node in another Group.

For example:

```
# ssh leader2 ctdb status
```

**8.** Ensure that no active `heal` operations are taking place at this time.

Log into one of the leader nodes, and run the `gluster volume heal` command. Proceed only if the output shows zeros. For example:

```
leader1:~ # gluster volume heal cm_logs info summary
Brick 172.23.0.3:/data/brick_cm_logs
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
.
.
.
```

**9.** From the admin node, run the `cm node update` command on each SU leader node in the Group.

Enter the following command:

```
# cm node update -n leader1,leader4,leader7 --repo-group hpcm16
```

**10.** Ensure that the cluster manager configuration files are compatible with the HPE Performance Cluster Manager 1.6 release.

Use the `pdsh` command, in the following format:

```
pdsh -w leaderX,leaderX,leaderX /opt/clmgr/lib/cluster-configuration
```

For *X*, specify the numbers of the leader nodes in the current Group.

For example:

```
# pdsh -w leader1,leader4,leader7 /opt/clmgr/lib/cluster-configuration
```

**11.** Reboot each SU leader node in the Group.

**12.** Check whether more Groups need to be upgraded.

To upgrade more Groups, return to Step **4**.

# 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 toolname enable
   cm monitoring toolname start
   ```

   For *toolname*, 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.

# Additional upgrade procedures

## Updating the software repository

Complete this procedure if you need to update the cluster software repositories at some time after you upgrade.

This topic explains how to update the software in the repositories. 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.

**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, complete the following steps to update the images for the 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 -o src_image_name -i image -r revision`

   The variables are as follows:

| Variable | Specification |
|---|---|
| *src_image_name* | The name of the source image. For example: `sles15sp3`. |
| *image* | A file name for the copied file (the clone). For example: `copy-sles15sp3`. |
| *revision* | A revision number. |

   Example 1. The following command backs up the compute node image:

   `# cm image copy -o rhel8.4 -i rhel8.4.backup`

   Example 2. The following command backs up the compute node image and tags the backup copy as a source-controlled copy. The command assumes that there are multiple versions of the source image that exist at this time. The command copies revision 2 of the source image to the backup.

   `# cm image copy -o rhel8.4 -i rhel8.4.backup -r 2`

5. Use the `cm image update` command to update the operating system software in each node image.

   For example, the following command updates the operating system image in a compute node image

   `# cm image update -i rhel8.4`

6. (Conditional) Create images for 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 Administration Guide**

   Obtain the operating system DVDs you need from the operating system vendor. Use cluster manager DVDs as needed.

## Upgrading AIOps without upgrading the cluster manager

The procedure in this topic explains how to upgrade the AIOps software. You can upgrade the AIOps software on a cluster anytime an AIOps software upgrade becomes available. You do not have to upgrade the cluster software itself as long as the upgrade is compatible with the cluster manager release on the cluster at this time.

### Prerequisites

The cluster manager hosts a software release level that is compatible with the AIOps software you want to install.

### Procedure

1. Retrieve the AIOps `.iso` file from the Hewlett Packard Enterprise customer portal.
2. Use the `cm repo add` command to add the AIOps software.

   For example:

   ```
   # cm repo add aiops-1.6-cd1-media-redhat-x86_64.iso
   ```

3. Use the `cm repo group add` command to add the cluster manager AIOps repository to a repository group.

   For example:

   ```
   # cm repo group add aiops-redhat-1.6 --repos Cluster-Manager-AIOps-1.6-rhel83-x86_64
   ```

4. Update the `aiops` packages.

   The following example command is for a RHEL 8.x platform:

   ```
   # cm node dnf -n admin update aiops-service aiops-config --repo-group aiops-redhat-1
   ```

5. Restart the `aiops` service:

   ```
   # cm aiops cooldev restart
   ```

# Websites

**General websites**

**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 initiates a fast and accurate resolution based on the service level of your product. 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.

**HPE Get Connected**

**https://www.hpe.com/services/getconnected**

**HPE Pointnext Tech Care**

**https://www.hpe.com/services/techcare**

**HPE Complete Care**

**https://www.hpe.com/services/completecare**

# 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, use the **Feedback** button and icons (located at the bottom of an opened document) on the Hewlett Packard Enterprise Support Center portal (**https://www.hpe.com/support/hpesc**) to send any errors, suggestions, or comments. All document information is captured by the process.

# YaST navigation

The following table shows SLES YaST navigation key sequences.

| Key | Action |
|---|---|
| **Tab**<br>**Alt** + **Tab**<br>**Esc** + **Tab**<br>**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**<br>**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*<br>**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. |

# Installing the operating system and the cluster manager separately

**Procedure**

1. **Preparing to install the operating system and the cluster manager separately**
2. **Installing and configuring the operating system**
3. **Installing the cluster manager**

## Preparing to install the operating system and the cluster manager separately

**Procedure**

1. Verify that this installation method can work for you by making sure that all of the following are true:

   - You want to install the operating system yourself so you can customize it.
   - You need 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.

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

   Select the nodes to act as leader nodes.

   The admin node needs access to the following:

   - Compute nodes
   - Node controllers (the 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 controllers.

3. 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. Make sure that the USB stick has a capacity of 16 GB or more.

  **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 of=/dev/sdX bs=512 count=65536
  # dd if=cm-admin-install-1.6-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**.

4. Plug the USB device into the admin node or mount the DVD.

# Installing and configuring the operating system

**Procedure**

1. Obtain the operating system installation software.

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

   **HPE Performance Cluster Manager operating system releases supported**

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

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

3. 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.

# 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-1.6-rhel8.4-x86_64.iso
   .
   .
   .
   # mount -o ro,loop cm-admin-install-1.6-rhel8.4-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. Respond to the prompts for the following information:

   a. The full path to the operating system distribution `.iso` file.

      This is the path that you used to install the admin node. The installer uses this information to set up repositories and start the installation.

   b. After the installer lists all the packages, the installer prompts you to enter **`y`** or **`n`** to proceed.

   c. After all packages are added, the installer prompts you to log out and log back in again.

   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 one of the following:

   - **Using the cluster definition file to specify the cluster configuration**

If you have a copy of the cluster definition file, use this procedure.

- **Using the menu-driven cluster configuration tool to specify the cluster configuration**

  If you do not have a copy of the cluster definition file, use this procedure. This procedure guides you through the menu-driven configuration tool.

# Upgrading the operating system and reinstalling the cluster manager

There are situations in which you might want to reinstall all or most of the software on a factory-configured cluster. A common case is the following:

- You are satisfied with the cluster configuration. That is, you want to reinstall the cluster system as it was configured at the factory with a minimum of changes.

  And

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

The procedure in this topic assumes that the cluster is intact and that you can back up the configuration files you need.

**Procedure**

1. **Backing up the configuration**
2. **Reinstalling the cluster manager**

## Backing up the configuration

**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
   ```

   The variables and parameters are as follows:

| Variable or parameter | Specification |
|---|---|
| `--images`<br>`--kernel`<br>`--kernel-parameters` | Specify these parameters if you plan to reinstall the same operating system images and kernel parameters. |
| `--ips` | Specify 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 . You can change these IP addresses later with the `cm node modify` command. |
| *cdf_backup_location* | A filename. |

   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. Enter the following commands to stop the cluster manager:

```
# systemctl stop config_manager.service
# systemctl stop clmgr-power.service
# systemctl stop cmdb.service
```

4. Enter the following command to back up the cluster database:

```
# 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"
```

5. Enter the following commands to start the cluster manager:

```
# systemctl start cmdb.service
# systemctl start clmgr-power.service
# systemctl start config_manager.service
```

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

7. (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 controllers.

---

**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.

---

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 mgmtsw`*X*

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

**9.** Enter the following command to display the repositories:

```
# cm repo show
```

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

**10.** Back up system images, system software repositories, and any other files you need to another system at your site.

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

# Reinstalling the cluster manager

**Procedure**

**1.** Install the cluster software on the admin node.

**a.** Complete the procedures in the following topics:

- **Installing the operating system and the cluster manager simultaneously on the admin node**
- **(Optional) Configuring a system admin controller high availability (SAC HA) admin node**. Complete this procedure if the admin node is an HA admin node.
- **Configuring the cluster software on the admin node**

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

**2.** (Conditional) Preserve the existing switch configuration.

Complete this step if you want to retain the current network, VLAN, and IP configuration in the cluster. That is, complete this step if you did not reset the management switches when you backed up the configuration.

Enter the following command to omit the switch configuration:

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

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

**3.** Run the `discover` command to configure the cluster.

**4.** (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 mgmtsw`*X*

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
```

5. (Conditional) Complete the scalable unit (SU) leader node configuration.

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

   Use the link at the end of this procedure under the heading **More Information**.

   **NOTE:** Return to this step in this procedure after you configure the SU leader nodes.

6. Direct the system to enable top-level switch configuration when you run a node discovery command in the future:

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

7. (Conditional) Recreate or import custom images, repositories, or files that you backed up.

   Complete this step as needed.

   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.

8. Enter the following command to reboot the cluster:

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

# 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 an HPE Cray EX cluster

HPE Cray EX clusters contain chassis management modules (CMMs), cabinet environment controllers (CECs), HPE Slingshot switch controllers, HPE Cray EX node controllers, and HPE Cray EX compute nodes.

By default, two networks are automatically generated for every eight CMMs (one HPE Cray EX logical cabinet) that are attached to a cluster. These two networks are separated by VLAN tagging at the CMM level in the network design, which matches the management switch port configuration for the appropriate VLAN settings. These two networks carry the following traffic:

- `hostmgmt`*`XXXX`* carries host management traffic such as compute node PXE, SSH, TFTP, and ICMP.
- `hostctrl`*`XXXX`* carries control traffic such as the CEC, CMMs. node controller, Redfish, SNMP, and so on.

By default, these networks use the following logic to determine VLAN numbers:

- Host management network (`hostmgmt`*`XXXX`*) = *cabinet_number* + 1000
- Host control network (`hostctrl`*`XXXX`*) = *cabinet_number* + 2000

For example, cabinet `1000` uses the following VLANs:

- VLAN 2000 for the host management network, with a network name of `hostmgmt2000`.
- VLAN 3000 for the host control network, with a network name of `hostctrl3000`.

The starting VLAN numbers used in HPE Cray EX clusters must be set to match the preceding logic. You can check and verify these settings with the `cadmin` command. For example:

```
# cadmin --show-mgmt-vlan-start
2000
# cadmin --show-mgmt-ctrl-vlan-start
3000
```

**NOTE:** The VLAN start and VLAN end numbers must match the HPE Cray EX physical hardware settings before you run the `cmcdetectd` command. This order ensures that the proper database, IP subnet, routing, and switch settings are applied. To set these VLAN numbers, use the following commands:

```
# cadmin --set-mgmt-vlan-start <#>
# cadmin --set-mgmt-ctrl-vlan-start <#>
```

Typically, there are two cabinet environment controllers (CECs) in an HPE Cray EX cabinet. These two controllers communicate to their CMMs to power on a cabinet. This communication occurs in the host control VLAN. Either configure the CECs in their proper host control VLAN by using the `switchconfig` command or by using a switch CLI. For example, by default, the two CECs in cabinet `1000` use VLAN 3000 for host control traffic. If these two CECs are plugged into switch `sw-cdu01` on ports 1/1/41 and 1/1/45, run the following command to assign these CECs to VLAN 3000:

```
# switchconfig set -s sw-cdu01 --default-vlan 3000 -p 1/1/41,1/1/45
```

# Network and subnet information within an HPE Apollo 9000 cluster

HPE Apollo 9000 clusters contain chassis controllers, HPE Adaptive Rack Cooling Systems (ARCS) components, or both. All components behind a chassis controller take on the network/VLAN settings of the chassis controller. An HPE Apollo 9000 chassis controller can have InfiniBand switches, compute nodes, CDUs, ARCS components, and PDUs behind it.

By default, two networks are automatically generated for every eight chassis controllers 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 node controller, Redfish, SNMP, IPMI, and power

If the automatic generation of networks is disabled, chassis controllers and their respective components end up in the `head` and `head-bmc` networks. To disable the automatic generation of networks, before you run a discovery command to configure any component, run the following command:

```
# cadmin --set-cmcs-per-mgmt-vlan 0
```

To verify the number of chassis controllers per management VLAN that is allowed, enter the following command:

```
# cadmin --show-cmcs-per-mgmt-vlan
```

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**.

# Network and subnet information within a cluster

**Table 7: 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 system-wide IP address ranges on the networks

Generally, a node and its node controller 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.

On systems such as the HPE Cray EX, the HPE Cray EX compute nodes use VLAN tagging to separate node controller traffic from node management traffic.

**Table 7: 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 7: Network and subnet information

| VLAN # | Subnet name | IP range / subnet mask | Nodes in Subnet |
|---|---|---|---|
| 1 | `head` | `172.23.0.0/16` | Admin `bond0` (untagged)<br>Generic compute (untagged)<br>Scalable unit (SU) leader (untagged)<br>Management switch VLAN 1 (untagged) |
| 1 | `head-bmc` | `172.24.0.0/16` | Admin `bond0:bmc` (untagged)<br>SU leader node controller (untagged)<br>Generic compute node controller (untagged)<br>Cooling devices such as ARCS (untagged)<br>Cooling devices such as CDUs (untagged) |
| 2001~2999<br>Example 1<br>Example 2 | `hostmgmt`*`XXXX`*<br>`hostmgmt2001`<br>`hostmgmt2002` | `10.168.0.0/13`<br>`10.168.0.0/22`<br>`10.168.4.0/22` | HPE Apollo 9000 compute (untagged) |
| 2001~2999<br>Example 1<br>Example 2 | `hostctrl`*`XXXX`*<br>`hostctrl2001`<br>`hostctrl2002` | `10.176.0.0/13`<br>`10.176.0.0/22`<br>`10.176.4.0/22` | HPE Apollo 9000 compute node controller (untagged)<br>HPE Apollo 9000 chassis controller (untagged)<br>HPE Apollo 9000 InfiniBand switch (untagged)<br>HPE Apollo 9000 CDU (untagged) |

*Table Continued*

| VLAN # | Subnet name | IP range / subnet mask | Nodes in Subnet |
|---|---|---|---|
| 2000~2999 | `hostmgmt`*`XXXX`* | `10.168.0.0/13` | HPE Cray EX compute (untagged) |
| Example 1 | `hostmgmt2000` | `10.168.0.0/22` | |
| Example 2 | `hostmgmt2001` | `10.168.4.0/22` | |
| 3000~3999 | `hostctrl`*`XXXX`* | `10.176.0.0/13` | HPE Cray EX cabinet management module (CMM) (tagged) |
| Example 1 | `hostctrl3000` | `10.176.0.0/22` | HPE Cray EX cabinet environment controller (untagged) |
| Example 2 | `hostctrl3001` | `10.176.4.0/22` | HPE Cray EX node controller (tagged) |
| | | | HPE Cray EX switch controller (tagged) |
| N/A | `ib0` | `10.148.0.0/16` | Any component with InfiniBand interfaces |
| N/A | `ib1` | `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 a node discovery 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 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 8: 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 8: Naming conventions**

| Component | Internal name format | Examples | Found in |
|---|---|---|---|
| Admin node | *admin* | `myadmin`<br>`sleet`<br>`snow` | All clusters |
| Ethernet management switch | `mgmtsw`*X* | `mgmtsw0` (spine)<br>`mgmtsw1` (leaf) | All clusters |
| Ethernet data switch | `dataswX` | `datasw0` (spine)<br>`datasw1` (leaf) | Clusters with an Ethernet high-speed fabric |
| InfiniBand data switch | `ibsw`*X* | `ibsw0`<br>`ibsw1` | Clusters with an InfiniBand high-speed fabric |
| Compute node | `service`*X* | `service1`<br>`service100` | Clusters with generic compute resources |
| Scalable unit (SU) leader node | `leader`*X* | `leader1`<br>`leader9` | Clusters with SU leader nodes |
| Node controller interfaces | `service`*X*`-bmc`<br>`leader`*X*`-bmc` | `service1-bmc`<br>`leader1-bmc` | Clusters that contain components that have a node controller. Node controllers can be of type iLO or BMC. |
| HPE Apollo 9000 chassis controller (nonadjustable) | `r`*X*`c`*Y* | `r1c1`<br>`r4c4` | Clusters with HPE Apollo 9000 hardware |
| HPE Apollo 9000 compute node | `r`*X*`c`*Y*`t`*T*`n`*Z* | `r1c1t2n1`<br>`r2c1t1n0` | Clusters with HPE Apollo 9000 hardware |
| Power distribution unit (PDU) | `pdu`*X* | `pdu0`<br>`pdu1` | Clusters with PDU hardware |

*Table Continued*

| Component | Internal name format | Examples | Found in |
|---|---|---|---|
| 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 |

# 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:

**(Optional) Configuring custom partitions on the admin node**

## Partition layout for a one-slot cluster

**Table 9: 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 9: 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 | XFS | `sgiroot` | Slot 1 `/` partition. |

## Partition layout for a two-slot cluster

**Table 10: Partition layout for a dual-boot cluster** shows the partition layout for a two-slot cluster. This layout yields two boot partitions.

**Table 10: 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.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 22 | VFAT | `sgiefi2` | Slot 2 `/boot/efi` partition.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 23-30 | N/A | N/A | N/A |
| 31 | XFS | `sgiroot` | Slot 1 `/` partition. |
| 32 | XFS | `sgiroot2` | Slot 2 `/` partition. |

# Partition layout for a five-slot cluster

**Table 11: Partition layout for a quintuple-boot cluster** shows the partition layout for a five-slot cluster. This layout yields five boot partitions.

**Table 11: 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.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 22 | VFAT | `sgiefi2` | Slot 2 `/boot/efi` partition.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 23 | VFAT | `sgiefi3` | Slot 3 `/boot/efi` partition.<br>EFI BIOS clusters only.<br>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.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 25 | VFAT | `sgiefi5` | Slot 5 `/boot/efi` partition.<br>EFI BIOS clusters only.<br>On x86_64 BIOS clusters, this partition is unused. |
| 26-30 | N/A | N/A | N/A |
| 31 | XFS | `sgiroot` | Slot 1 / partition. |
| 32 | XFS | `sgiroot2` | Slot 2 / partition. |
| 33 | XFS | `sgiroot3` | Slot 3 / partition. |
| 34 | XFS | `sgiroot4` | Slot 4 / partition. |
| 35 | XFS | `sgiroot5` | Slot 5 / partition. |

# Specifying configuration attributes

The cluster manager include the following types of configuration attributes:

- Node attributes
- Cluster attributes

The cluster definition file includes all the node attributes and cluster attributes assigned in the cluster. In the cluster definition file, node attributes are found in the following sections:

- `[discover]`
- `[nic_templates]`
- `[templates]`

In the cluster definition file, cluster attributes are found in the following sections:

- `[attributes]`
- `[dns]`
- `[images]`
- `[networks]`

To obtain a copy of the cluster definition file, enter the following command:

```
# discover --show-configfile
```

The cluster manager commands and utilities use the configuration attributes to define individual nodes and to define the general cluster in the following way:

- The `configure-cluster` command, starts a menu-driven utility that you can use at any time to specify cluster attributes. The command is as follows:

  ```
  # configure-cluster
  ```

  Alternatively, you can provide a cluster definition file, populated with attributes, as input to the `configure-cluster` command. The format is as follows:

  ```
  configure-cluster --configfile cluster_definition file
  ```

- The `cm node add` command and the `cm node discover` command assign node attributes to nodes when they configure nodes into the cluster.

---

**NOTE:** In many cases, you can set or clear attributes on a command line. On a command line, the attribute name often uses underscore characters (_). In the cluster definition file, the attribute name often uses hyphens (–). For example, you can set the UDPcast attribute `udpcast_max_bitrate` in the cluster definition file. However, on the `cm node set` command line, 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
- `cm node show` command
- `cm node set` command
- `cm node show` command
- `discover` command

## transport

Sets the image transport method.

Values = `rsync`, `bt`, `udpcast`

Default = `bt`

Range = N/A

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

# UDPcast options

## admin_udpcast_mcast_rdv_addr

When UDPcast is used, this attribute specifies the UDPcast rendezvous (RDV) address at which admin node senders and non-admin receiver nodes find each other.

The receiver nodes can be scalable unit (SU) leader nodes or compute nodes.

Default = `239.255.255.1`.

Values = any valid IPv4 multicast address in the 224.0.0.0/4 range .

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cm node show` command

  To display the cluster-wide setting for this attribute, enter the following command:

  `# cm node show --udpcast-mcast-rdv-addr -n admin`

  To display a node-specific setting for this attribute, enter the `cm node show` command in the following format:

  `cm node show --udpcast-mcast-rdv-addr -n` *`node`*

  For *node*, specify the node hostname.
- `cattr` command
- `cm node set` command

  To set a node-specific value for this attribute, use the `cm node set` command in the following format:

  `cm node set --udpcast-mcast-rdv-addr` *`value`* `-n` *`node`*

  For *value*, specify the node-specific value.

  For *node*, specify the node hostname.

  If you specify a nondefault IP address, also use the `cm node provision` command to push an image and initiate changes on the receiver nodes.
- `discover` command

## 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
- `cm node show` 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
- `cm node show` 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
- `cm node show` 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
- `cm node show` 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.

For the admin node, the default is `5000` (5 seconds).

For the leader nodes, the default is `0`.

Values = any positive integer number. When set to `0`, this attribute is disabled.

Accepted by:

- Cluster definition file
- `cm node show` 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, the cluster manager sets the `udpcast_rexmit_hello_interval` value to 5000 (5 seconds) for UDP senders running on the admin node. Enter the following command to display this value:

```
# cm node show --udpcast-rexmit-hello-interval --global
```

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

The following command displays the attribute value for SU leader node `leader1`:

```
# cm node show --udpcast-rexmit-hello-interval -n leader1
```

The admin node uses the global value when it serves leader nodes and compute nodes using the UDPcast transport mechanism.

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.

For the admin node, the default is `2`. The admin nodes serve the leader nodes and the 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`
- `cm node show` command
- `discover` command

## Management network subnet and VLAN attributes

The topics that follow provide information about the configuration attributes that the cluster manager uses for the management network on HPE Cray EX clusters and HPE Apollo 9000 clusters.

The attribute settings appear in the cluster definition file. You can use the `cadmin` command to display or to change the VLAN attributes. For example:

```
# cadmin --show-mgmt-vlan-start
.
```

```
.
.
# cadmin --set-mgmt-vlan-start
.
```

**NOTE:** The installation process guides you through configuring the VLAN and running the node discovery commands. That is the correct order. Apply the VLAN settings before any nodes are configured into the cluster and before the `cmcdetectd` service starts. In the menu-driven cluster configuration tool, the path to access these settings is as follows:

**Initial Setup Menu > Network Settings > Configure Management Network VLAN Settings**

Hewlett Packard Enterprise strongly recommends that if you need to change any management network settings, that you change the settings before you run the node discovery commands to configure any cluster components and before the `cmcdetectd` or `cmcinventory` services start. It requires a full database reset to adjust these settings if nodes are already in the database.

## cmcs_per_mgmt_vlan

The cluster manager supports this configuration attribute on HPE Cray EX systems and HPE Apollo 9000 systems only.

Specifies the number of chassis controllers included in one management VLAN. When the number of chassis controllers in one VLAN reaches that number, the cluster manager generates an additional pair of subnets. A subnet includes a management VLAN and a control VLAN.

Default = `8`.

Values = must be a multiple of `4`. For example, `4`, `8`, `16`.

Range = `0` <= *vlan* <= `48`, as follows:

- Larger *vlan* values result in a cluster with larger broadcast domains per VLAN. These larger VLANs include more compute nodes and more hardware.
- When *vlan* is set to `0`, the cluster manager disables routed management networking completely. To disable this feature, set this value to `0` before any management switches are configured into the cluster.

  When disabled, the `cmcdetectd` and `cmcinventory` services puts all hardware configured automatically into the `head` and `head-bmc` networks on VLAN 1. Hewlett Packard Enterprise does not recommend disabling this feature. When disabled, broadcast domains quickly become excessively large for clusters with multiple hardware racks.

Accepted by:

- Cluster configuration tool
- `cadmin` command
- `cattr` command

## cmcs_per_rack

The cluster manager supports this configuration attribute on HPE Cray EX systems and HPE Apollo 9000 systems only.

This attribute specifies the number of chassis controllers per hardware rack.

Default = `4`.

Accepted by:

- Cluster configuration tool
- `cadmin` command
- `cattr` command

## cmms_per_rack

The cluster manager supports this configuration attribute on HPE Cray EX systems only.

This attribute specifies the number of chassis management module (CMM) controllers per hardware rack.

Default = `8`.

Accepted by:

- Cluster configuration tool
- `cadmin` command
- `cattr` command

## mgmt_ctrl_vlan_end

The cluster manager supports this configuration attribute on HPE Cray EX systems only.

This attribute specifies the last VLAN number for the management subnet used for node controller traffic. No VLANs are used past this setting. This setting specifies the upper limit on the number of VLANs allowed to be created.

Default = `3999`.

Range = `2` <= *vlan* <= `4091`. The *vlan* number must be higher than the value for `mgmt_ctrl_vlan_start`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

## mgmt_ctrl_vlan_start

The cluster manager supports this configuration attribute on HPE Cray EX systems only.

The cluster manager uses this setting only where the control traffic resides, which occurs between the chassis environment controller (CEC) and the chassis management module (CMM). The VLAN configured on the Layer 3 switch for a given HPE Cray EX rack must match the CEC front panel setting for the VID (VLAN ID).

This attribute specifies the first VLAN number for the management subnet used for node controller traffic. Each subsequent subnet is assigned to the next incremental *vlan* + 1.

Default = `3001`.

Range = `2` <= *vlan* <= `4091`. The *vlan* number must be lower than the value for `mgmt_ctrl_vlan_end`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

## mgmt_net_subnet_selection

The cluster manager supports this configuration attribute on HPE Cray EX and HPE Apollo 9000 systems only.

The `mgmt_net_subnet_selection` attribute specifies how subnets are assigned to racks.

Default = `mgmt_net_subnet_selection=rack-based` (default) or `mgmt_net_subnet_selection=next-available`

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cattr` command

The attribute settings are as follows:

- `mgmt_net_subnet_selection=rack-based` (default).

  Hewlett Packard Enterprise recommends that you use the default setting when possible.

  The `cmcdetectd` service uses the value of the `rack_start_number` attribute in conjunction with the `mgmt_net_subnet_selection=rack-based` attribute for the following operations:

  - When determining how many racks away a given chassis controller is located.
  - When assigning subnets. When the rack start number is set and `mgmt_net_subnet_selection=rack-based`, you can power-on the hardware in any order and maintain sequential subnet assignments.

- `mgmt_net_subnet_selection=next-available`. Specify only if the cluster rack numbering is non-sequential. For example, if racks are numbered 1 through 10 and then 3000 to 3010.

  With this selection scheme, the cluster manager assigns subnets on a first-come-first-served basis. The assignments can appear random if all hardware is powered on at the same time before the database is fully configured.

Example 1. Assume you have an HPE Apollo 9000 cluster with the following characteristics:

- Six racks of compute nodes.
- Four chassis controllers per rack.
- Rack numbers start at `1` and go up through `6`.

  The rack topology is as follows:

  - Rack 1 contains four chassis controllers: `r1c1`, `r1c2`, `r1c3`, and `r1c4`.
  - Rack 2 contains four chassis controllers: `r2c1`, `r2c2`, `r2c3`, and `r2c4`.

- Rack 3 contains four chassis controllers: `r3c1`, `r3c2`, `r3c3`, and `r3c4`.
- Rack 4 contains four chassis controllers: `r4c1`, `r4c2`, `r4c3`, and `r4c4`.
- Rack 5 contains four chassis controllers: `r5c1`, `r5c2`, `r5c3`, and `r5c4`.
- Rack 6 contains four chassis controllers: `r6c1`, `r6c2`, `r6c3`, and `r6c4`.

- There are eight chassis controllers per management VLAN, and the starting management VLAN is numbered `2001`.

The cluster manager generates the following subnets:

- Racks `1` and `2` (VLAN `2001`):
  - `hostmgmt2001 = 10.168.0.0/22`
  - `hostctrl2001 = 10.176.0.0/22`
- Racks `3` and `4` (VLAN `2002`):
  - `hostmgmt2002 = 10.168.4.0/22`
  - `hostctrl2002 = 10.176.4.0/22`
- Racks `5` and `6` (VLAN `2003`):
  - `hostmgmt2003 = 10.168.8.0/22`
  - `hostctrl2003 = 10.176.8.0/22`

Example 2. Assume that you have an HPE Cray EX cluster with the following characteristics:

- Four cabinets of compute nodes.
- A cluster definition file that defines the following attributes:

```
mgmt_vlan_start           : 2000
mgmt_vlan_end             : 2999
mgmt_ctrl_vlan_start      : 3000
mgmt_ctrl_vlan_end        : 3999
cmcs_per_mgmt_vlan        : 8
cmcs_per_rack             : 8
rack_start_number         : 1000
mgmt_net_subnet_selection : rack-based
```

- Eight chassis controllers per rack. Rack numbers start at `1000` and go up through `1003`.

  The rack topology is as follows:
  - Rack `1000` contains eight chassis controllers: x1000c0~x1000c7
  - Rack `1001` contains eight chassis controllers: x1001c0~x1001c7
  - Rack `1002` contains eight chassis controllers: x1002c0~x1002c7
  - Rack `1003` contains eight chassis controllers: x1003c0~x1003c7

- Eight chassis controllers per management VLAN. The starting management VLAN is `2000`. The starting management control VLAN is `3000`.

The cluster manager generates the following subnets:

- Rack `1000` (VLAN `2000` and VLAN `3000`):
  - `hostmgmt2000 = 10.168.0.0/22`
  - `hostctrl3000 = 10.176.0.0/22`
- Rack `1001` (VLAN `2001` and VLAN `3001`)
  - `hostmgmt2001 = 10.168.4.0/22`
  - `hostctrl3001 = 10.176.4.0/22`
- Rack `1002` (VLAN `2002` and VLAN `3002`)
  - `hostmgmt2002 = 10.168.8.0/22`
  - `hostctrl3002 = 10.176.8.0/22`
- Rack `1003` (VLAN `2003` and VLAN `3003`)
  - `hostmgmt2003 = 10.168.12.0/22`
  - `hostctrl3003 = 10.176.12.0/22`

## mgmt_vlan_end

The cluster manager supports this configuration attribute on HPE Cray EX platforms and HPE Apollo 9000 platforms only.

On HPE Cray EX platforms, management and control subnets are assigned to different VLAN numbers.

On HPE Apollo 9000 platforms, the cluster manager uses the management VLAN for both management traffic and control (iLO) traffic despite using different subnet ranges. To manage these tasks, the cluster manager assigns two IP addresses to one VLAN and uses an Access Control List (ACL) on the Ethernet switch to block traffic.

This attribute is the end point of the `mgmt_vlan_start` attribute. It specifies the last compute node rack VLAN. No VLANs are used past this setting. This setting specifies the upper limit on the number of VLANs that can be created.

Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2999`.

Range = `2` <= *vlan* <= `4091`. The *vlan* number must be higher than the value for `mgmt_vlan_start`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

## mgmt_vlan_start

The cluster manager supports this configuration attribute on HPE Cray EX platforms and HPE Apollo 9000 platforms only.

On HPE Cray EX platforms, management and control subnets are assigned to different VLAN numbers.

On HPE Apollo 9000 platforms, the cluster manager uses the management VLAN for both management traffic and control (iLO) traffic despite using different subnet ranges. To manage these tasks, the cluster manager assigns two IP addresses to one VLAN and uses an Access Control List (ACL) on the Ethernet switch to block traffic.

This attribute specifies the first compute node rack VLAN. This VLAN is the first management subnet used for regular compute traffic. Each subsequent subnet is assigned to the next incremental *vlan* + 1.

Use caution when changing this value. Take care not to overlap other VLAN settings.

Default = `2001`.

Range = `2` <= *vlan* <= `4091`. The *vlan* number must be lower than the value for `mgmt_vlan_end`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cattr` command
- `cadmin` command

## rack_start_number

The cluster manager supports this configuration attribute on HPE Cray EX platforms and HPE Apollo 9000 platforms only.

This attribute specifies the starting number for the racks in the cluster.

Default = `1`

Set this variable to the number of the first physical rack in the cluster.

The `cmcdetectd` service uses the value of the `rack_start_number` attribute in conjunction with the `mgmt_net_subnet_selection=rack-based` attribute for the following operations:

- When determining how many racks away a given chassis controller is located.
- When assigning subnets. When the rack start number is set and `mgmt_net_subnet_selection=rack-based`, you can power-on the hardware in any order and maintain sequential subnet assignments.

Accepted by:

- Cluster configuration tool
- `cadmin` command
- `cattr` command

## redundant_mgmt_network

Specifies the default setting for the redundant management network. If no value is supplied at configuration time, the installer populates all nodes with the default value.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cm node show` command
- `cattr` command
- `cm node set` command
- `discover` command

### switch_mgmt_network

Specifies the default setting for the switch management network. If no value is supplied at configuration time, the installer populates all nodes with the default value.

Values = `yes` (default) or `no`.

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cm node show` command
- `cattr` command
- `cm node set` command
- `discover` command

## Console server options

The admin node and the leader nodes are management nodes.

On a management node, there are files in the `/var/log/consoles` directory for each subordinate node. The files contain log information from the node controllers on the subordinate nodes.

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 node controller 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
- `cm node show` 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
- `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 interfaces in Linux.

By default, `eth0` and `eth1` are used on leader nodes and on compute nodes.

For more information about predictable interface naming, see the documentation for your operating system. Generally, different types of compute hardware and NICs have different naming styles. Common formats include the following:

- `eno#`
- `ens#f#`
- `enp#f#`

If you specify more than one address, 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 = At least one, and up to 64 interface hostnames. If you specify more than one interface, use a comma (`,`) to separate 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.

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 configuring 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
- `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
- `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
- `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
- `cm monitoring native` command

## Miscellaneous options

### alias_groups

Defines node aliases, which are additional ways to refer to a node, at cluster configuration time. This attribute adds the `alias_groups` keyword within the node definitions of the `[discover]` section of the cluster definition file.

The format is as follows:

`alias_groups="`*`group1`*`:`*`alias1`*`,`*`group2`*`:`*`alias2`*`,..."`

The variables are as follows:

| Variable | Specification |
|---|---|
| *groupX* | The name for the group of nodes. |
| *aliasX* | The name for the individual nodes. |

Example:

In the `[discover]` section, a node with `hostname1=node1` could define an aliases called `work1` in the alias group `work-nodes` by adding the following to the configuration definition file:

```
alias_groups="work-nodes:work1"
```

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cm node set` command
- `cm node show` command
- `discover` command

## 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

## baud_rate

Specifies the baud rate of the serial console device.

Values = a positive integer value

Default = `115200`

**NOTE:** This attribute is required.

Range = NA

Accepted by:

- Cluster definition file
- `cm node set` command
- `cm node show` command
- `discover` command

## bmc_password

Specifies the node controller password.

Values = a cluster-specific node controller password.

**NOTE:** This attribute is case sensitive and is required.

Accepted by:

- Cluster definition file
- `cm node set` command
- `cm node show` command
- `discover` command

## bmc_username

Specifies the node controller username.

Values = a cluster-specific node controller username.

**NOTE:** This attribute is case sensitive and is required.

Accepted by:

- Cluster definition file
- `cm node set` command
- `cm node show` command
- `discover` command

## card_type

Specifies the type of node controller in the node.

Values = Valid values include `IPMI` (default), `bmx`, `iLO`, and `ILOCM`. Specifically, 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.

Default = `IPMI`.

**NOTE:** This attribute is case sensitive and is required.

Range = NA

Accepted by:

- Cluster definition file
- `cm node set` command

- `cm node show` command
- `discover` command

## cluster_domain

This configuration attribute 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`

## custom_partitions

The custom partitioning feature is not enabled by default. The cluster manager supports this feature on nodes with root disks.

The `custom_partitions` variable specifies the name of the custom partition file. The file resides in the following directory:

`/opt/clmgr/image/scripts/pre-install`

Values = any file name that ends in `.cfg`

Accepted by:

- Cluster configuration tool
- Cluster definition file
- `cadmin` command
- `cm node set` command
- `cm node show` command
- `discover` command

## dhcp_bootfile

This attribute specifies whether to load iPXE (the default) or GRUB2 first when a node is first configured.

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.

The settings are as follows:

- The default boot loader is `dhcp_bootfile=ipxe-direct`. In this case, the server boot agent uses a special iPXE binary on UEFI and legacy BIOS systems to directly load the kernel and `initrd`. It avoids GRUB2.

HPE Cray EX compute nodes and HPE Apollo 20 compute nodes require `ipxe-direct`.

- When you specify `dhcp_bootfile=grub2`, GRUB2 loads first.

  The `grub2` specification is the only specification supported on Arm (AArch64) platforms.

Use the `cm node set` command to specify the new boot order.

Values = `grub2` (default), `ipxe-direct`, `ipxe`.

Accepted by:

- Cluster definition file
- `cadmin` command
- `cm node show` command
- `cm node set` command
- `discover` command

The following other topics might also interest you:

- **`dhcpd_default_lease_time`**
- **`dhcpd_max_lease_time`**
- **Nodes fail to boot**

## dhcpd_default_lease_time

This attribute specifies the default DHCP lease time, which the cluster manager sets at 180 seconds.

The cluster manager includes other configuration attributes that let you control DHCP, but Hewlett Packard Enterprise recommends that you use the default values whenever possible.

You can also use the `cadmin` command to set this default lease time.

Values = `180` seconds (default) or another integer value.

Accepted by:

- Cluster definition file
- `cadmin` command

The following topics might also interest you:

- **`dhcp_bootfile`**
- **`dhcpd_max_lease_time`**
- **Nodes fail to boot**

## dhcpd_max_lease_time

This attribute specifies the maximum DHCP lease time, which the cluster manager sets at 300 seconds.

The cluster manager includes other configuration attributes that let you control DHCP, but Hewlett Packard Enterprise recommends that you use the default values whenever possible.

You can also use the `cadmin` command to set this maximum lease time.

Values = `300` seconds (default) or another integer value.

Accepted by:

- Cluster definition file
- `cadmin` command

The following topics might also interest you:

- **dhcp_bootfile**
- **dhcpd_default_lease_time**
- **Nodes fail to boot**

## 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`. If the node uses an NFS root file system or a `tmpfs` root file system, specify `disk_bootloader=no`.

Default = `no`

Range = NA

Accepted by:

- Cluster definition file
- `cm node show` 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
- `cm node show` command
- `cattr` command
- `discover` command

## geolocation

Specifies information about the physical location of the node at your site in a human-readable form.

Values = An alphanumeric string of up to 128 characters. The string can include spaces and special characters. If you include spaces, enclose the string in quotation mark (") characters.

For more information, see the following:

**(Conditional) Configuring power distribution units (PDUs) into the cluster**

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `cm node show` 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

## 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

## kernel_distro_params

Specifies kernel parameters for the operating system distribution that runs on the cluster. Typically, this setting includes parameters suggested by the distribution.

You can set this parameter on either a node basis or an image basis. The node setting overrides the image setting.

Values = `arg=value,arg=value,...`

Accepted by:

- `cadmin` command
- `cattr` command
- `cm node set` command
- `cm node show` command
- `discover` command

For example:

```
# cm node set --kernel-distro-params cgroup_disable=memory,max_cstate=1 --nodes service1
For ICE compute nodes, the image must be pushed to the leaders to take effect.
Refreshing netboot files for impacted nodes...

# cm node show --kernel-distro-params --nodes service1
cgroup_disable=memory,max_cstate=1

# cm node unset --kernel-distro-params --nodes service1
For ICE compute nodes, the image must be pushed to the leaders to take effect.
Refreshing netboot files for flat compute, leader nodes...

# cm node show --kernel-distro-params --nodes service1
None
```

For information about specifying parameters on a command line, see the help output (`-h`) for a specific command.

## kernel_extra_params

Specifies additional kernel parameters for the operating system distribution that runs on the cluster. This attribute sets parameters in addition to the standard parameters suggested by the distribution.

You can set this parameter on either a node basis or an image basis. The node setting overrides the image setting.

Values = `arg=value,arg=value,...`

Accepted by:

- `cadmin` command
- `cattr` command
- `cm node set` command
- `cm node show` command
- `discover` command

For example:

```
# cm node set --kernel-extra-params cgroup_disable=memory,rescue-1 --nodes service1
For ICE compute nodes, the image must be pushed to the leaders to take effect.
Refreshing netboot files for impacted nodes...

# cm node show --kernel-extra-params --nodes service1
cgroup_disable=memory,rescue-1

# cm node unset --kernel-extra-params --nodes service1
For ICE compute nodes, the image must be pushed to the leaders to take effect.
Refreshing netboot files for flat compute, leader nodes...

# cm node show --kernel-extra-params --nodes service1
None
```

For information about specifying parameters on a command line, see the help output (`-h`) for a specific 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

## node_notes

Specifies node-specific information in a human-readable form. You can include any information about the node in the note.

Values = An alphanumeric string of up to 8192 characters.

Default = NA

Range = NA

Accepted by:

- `cadmin` command
- `cattr` command
- `cm node set` command
- `cm node show` command

## pdu_protocol

This attribute specifies the protocol that the power distribution unit (PDU) uses.

Values = `SNMP` or `MODBUS`.

Default = `SNMP`

For more information, see the following:

**(Conditional) Configuring power distribution units (PDUs) into the cluster**

Accepted by:

- Cluster definition file
- `cadmin` command
- `cattr` command
- `cm node set` command
- `cm node show` command
- `discover` command

## 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
- `cm node show` 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 compute node is attached.

Values = the IP address of an SU leader node

Default = NA

Range = NA

Accepted by:

- Cluster definition file
- `cm node show` 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

# 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.

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. The cluster manager 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 12: 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` |
| C1104-TY13 compute | Management network | `enp1s0f0` | `eth0` |
| | House network | `enp1s0f1` | `eth1` |

The cluster manager includes the following commands for adding, deleting, modifying, and displaying NIC information:

- `cm node nic add`
- `cm node nic delete`
- `cm node nic set`
- `cm node nic show`

For information about these commands, enter the command and add the `-h` parameter. For example:

```
# cm node nic add -h
```

# Managing node additions and deletions on large cluster systems

On cluster systems with thousands of nodes, the `cm node add` command, the `cm node discover` command, and 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 the `fastdiscover` command.

**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 on the admin node:

   ```
   # cm node update config --sync -n admin
   ```
5. (Conditional) Update the internal configuration files on the scalable unit (SU) leader nodes.

   Complete this step on clusters with SU leader nodes.

   Enter the following command:

   ```
   # cm node update config --sync -t role su-leader
   ```
6. 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`
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**

# Configuring a new switch

New switches require some preliminary configuration before you configure them into a cluster. After you complete the preliminary configuration, you can run a node discovery 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 HPE Performance Cluster Manager Release Notes list the switches that are supported.

---

**NOTE:** To replace an existing switch for which you have backup data, use the procedure in the following:

**HPE Performance Cluster Administration Guide**

---

**Procedure**

1. **(Conditional) Configuring an Extreme Networks switch**
2. **(Conditional) Configuring an HPE FlexFabric switch or an HPE FlexNetwork switch**
3. **Running the `cm node add` command for a new switch**

## (Conditional) Configuring 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"
```

# (Conditional) Configuring 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"
   ```

# Running the `cm node add` 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 a cluster definition file example for a spine switch:

   ```
   # Spine Switch
   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 a cluster definition file example for a leaf switch:

   ```
   # Leaf Switch
   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
   ```

   **NOTE:** If the cluster does not have HPE SGI 8600 ICE hardware, set the `ice=` configuration attribute to `no`. Example: `ice=no`.

   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 simultaneously on the admin node**

   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 `cm node add` command in the following format:

   To configure one management switch, use the following format:

   `cm node add -c` *`path_to_configfile`*

   For *path_to_configfile*, specify the full path to the cluster definition file.

   For example:

   ```
   # cm node add -c mgmtsw.config
   Config file: mgmtsw.config
   Add - All nodes in the mgmtsw.config will be added to the database.
   hog1: fastdiscover: Config file parse step: , 0.06s
   hog1: fastdiscover: new nodes step: , 0.09s
   hog1: fastdiscover: Script time: , 0.16s

   Refreshing the netboot environment for nodes in the config file...

   Updating admin node configs...
   Configuration manager initiating node configuration.
   0 of 1 nodes completed in 34.0 seconds, averaging 0.0s per node
   1 of 1 nodes completed in 34.8 seconds, averaging 34.1s per node
   Node configuration complete.

   Updating SU leader configs...
   Configuration manager initiating node configuration.
   0 of 3 nodes completed in 7.1 seconds, averaging 0.0s per node
   3 of 3 nodes completed in 7.6 seconds, averaging 6.8s per node
   Node configuration complete.

   Performing switch configuration...

   Please view '/var/log/switchconfig.log' to verify no switch configuration error occurred during this process.
   ```

4. (Optional) Monitor the `cm node add` command progress.

After the `cm node add` process is complete, it takes time for the switches to obtain an IP address via DHCP and become fully configured.

To monitor the switch configuration progress, enter the following command:

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

After the `cm node add` command completes successfully, optionally continue to the next step.

5. (Optional) Enter the following command to verify that the firmware version matches this installation of the cluster manager:

   ```
   switchconfig sanity_check -s mgmtswX | grep firmware
   ```

   For *X*, specify the management switch number as it appears in the cluster definition file.

   Example 1: The following command is for an Extreme Networks switch:

   ```
   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: The following command is for an HPE switch:

   ```
   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')
   ```

6. (Conditional) Upgrade or downgrade the firmware to match the recommended release.

   Complete this step if the switch firmware does not match the firmware version recommended by the cluster manager.

   The cluster manager has support for some brands of switches to simplify the upgrade or downgrade process. Enter the following command and view the help output for more information:

   ```
   # switchconfig update_firmware --help
   ```

   For example, assume that you need to update `mgmtsw0`, which is an HPE FlexFabric 5940 48SFP + 6QSFP + switch. You want to update the switch to firmware version 7.1.070 Release 2612P08-US. Enter the following command:

   ```
   # switchconfig update_firmware --switches mgmtsw0 --update --firmware file 5940-CMW710-R2612P08-US.ipe
   ```

   The `switchconfig` command in this example assumes that file `5940-CMW710-R2612P08-US.ipe` resides in the default TFTP directory, which is `/opt/clmgr/tftpboot/`. To guide you through the firmware upgrade or downgrade process, the command prompts you to answer a series of questions. Optionally, add the `--force` option to suppress the prompts and just upgrade the switch firmware.

   The cluster manager supports the `switchconfig update_firmware` command on the following switches:

   - HPE FlexNetwork switches
   - HPE FlexFabric switches
   - HPE Aruba switches

# Configuring a serial console

If the system console you typically use is a graphical console, you can configure a serial console to make remote maintenance easier.

**Procedure**

1. Use a text editor to open file `/etc/default/grub`.
2. On the `GRUB_CMDLINE_LINUX_DEFAULT` line, edit the line to include `console=`*`settings`* and remove `splash=silent quiet`.

   For *settings*, specify values that match your system settings. For example:

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

3. On the `GRUB_TERMINAL` line, edit the line to change `gfxterm` to `console`.

   For example:

   ```
   GRUB_TERMINAL="console"
   ```

4. 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, if you want to remove the serial setup and switch back to the graphical console by default, complete the following steps:

   a. From the `GRUB_CMDLINE_LINUX_DEFAULT=` line, remove all the `console=` parameters.

   b. Change the `GRUB_TERMINAL` line back to `GRUB_TERMINAL="gfxterm"`