



SMW HA Installation Guide (S-0044-F)

Contents

About the SMW High Availability Installation Guide.....	5
Distribution Media.....	6
SMW HA Overview.....	7
SMW Cluster Configuration.....	7
Shared Storage on the Boot RAID.....	8
Storage for the Power Management Database (PMDB).....	9
Synchronized Files.....	9
Add Site-specific Files to the Synchronization List.....	10
Cluster Resources.....	11
Limitations of SMW Failover.....	12
Install a Cray SMW HA System.....	14
Prepare to Install a New SMW HA System.....	14
Shared Storage for SMW HA.....	15
Network Connections for an SMW HA System.....	16
Configuration Values for an SMW HA System.....	16
Passwords For an SMW HA System.....	18
Install SMW Software on the First SMW.....	18
Prepare to Install SMW Software on the First SMW.....	19
Install the SMW Release Package.....	19
Finish Installing SMW Software on the First SMW.....	46
Configure the Boot RAID for SMW HA.....	47
Configure the Boot RAID.....	47
Determine the Persistent Device Name for a LUN.....	65
Install CLE Software on the First SMW.....	65
Prepare to Install a New System.....	66
Configure the Boot RAID.....	67
About Installation Configuration Files.....	74
Install CLE on a New System.....	91
Install SMW Software on the Second SMW.....	141
Prepare to Install SMW Software on the Second SMW.....	142
Install the SMW Release Package.....	142
Finish Installing SMW Software on the Second SMW.....	156
Install CLE Software on the Second SMW.....	157
Install SMW HA Software.....	158
Verify the SMW and CLE Configuration.....	158

Install the SMW HA Release Package on Both SMWs.....	160
Configure the Cluster.....	161
Configure Required Cluster Settings.....	161
Configure Boot Image Synchronization.....	169
Configure Failover Notification.....	171
Configure PMDB Storage.....	172
Verify the SMW HA Cluster Configuration.....	181
Back Up a Newly-installed SMW HA System.....	182
R815 SMW: Create an SMW Bootable Backup Drive.....	182
(Optional) R815 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device.....	190
R630 SMW: Create an SMW Bootable Backup Drive.....	192
(Optional) R630 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device.....	199
Change Default HA Passwords After Installation.....	201
Move Local Log Data to RAID.....	202
Customize a Preinstalled SMW HA System.....	204
Prepare to Customize an SMW HA System.....	204
Customize the First SMW.....	205
Change the Cluster Configuration on the First SMW.....	206
Customize the Second SMW.....	208
Finish Customizing a Preinstalled SMW HA System.....	209
Verify Cluster Status After Customization.....	210
Back Up a Customized SMW HA System.....	211
R815 SMW: Create an SMW Bootable Backup Drive.....	211
(Optional) R815 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device.....	219
R630 SMW: Create an SMW Bootable Backup Drive.....	221
(Optional) R630 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device.....	228
Change Default Passwords After Customization.....	230
Configure PMDB Storage.....	231
Configure Mirrored Storage with DRBD for the PMDB.....	231
Configure Shared Storage on the Boot RAID for the PMDB.....	236
Upgrade the Cray SMW HA System.....	241
Before You Start an SMW HA Upgrade.....	241
Upgrade the Operating System Software.....	242
Upgrade the SMW Software.....	243
Prepare the SMW HA System for an SMW Upgrade.....	244
Upgrade SMW Software on the Active SMW.....	245
Upgrade SMW Software on the Passive SMW.....	262
Finish the SMW Upgrade.....	282

Upgrade SMW HA Software	285
Upgrade CLE Software on the SMW HA System.....	291
Upgrade CLE Software on the Active SMW.....	292
Copy Boot Images and CLE Install Directory to the Passive SMW.....	311
Upgrade CLE Software on the Passive SMW.....	311
Finish the CLE Upgrade.....	313
Configure PMDB Storage.....	313
Configure Mirrored Storage with DRBD for the PMDB.....	313
Configure Shared Storage on the Boot RAID for the PMDB.....	319
Migrate PMDB Data from the Boot RAID to Mirrored Storage.....	322
Update the Cray SMW HA System.....	325
Before You Start an SMW HA Update.....	325
Update SMW Software.....	326
Prepare the SMW HA System for an SMW Update.....	326
Update SMW software on the Active SMW.....	328
Update SMW Software on the Passive SMW.....	344
Finish the SMW Update.....	355
Update SMW HA Software.....	358
Update CLE Software on the SMW HA System.....	365
Upgrade CLE Software on the Active SMW.....	365
Copy Boot Images and CLE Install Directory to the Passive SMW.....	385
Upgrade CLE Software on the Passive SMW.....	385
Finish the CLE Upgrade.....	386
Configure PMDB Storage.....	387
Configure Mirrored Storage with DRBD for the PMDB.....	387
Configure Shared Storage on the Boot RAID for the PMDB.....	392
Migrate PMDB Data from the Boot RAID to Mirrored Storage.....	396
Troubleshooting SMW HA Installation Problems.....	399
Restore a Previous SMW HA Configuration.....	399
Disable the SMW HA Configuration.....	400
Re-enable the SMW HA Configuration.....	403
Migrate PMDB Data from Mirrored Storage to the Boot RAID.....	406
Remove the Mirrored Storage Disk for the PMDB.....	408

About the SMW High Availability Installation Guide

This publication provides software installation procedures for setting up an SMW HA system (also called SMW Failover, or SMW cluster). An SMW HA system is a Cray XC system with two second-generation rack-mount SMWs, either Dell R815 or Dell 630 models. The SMWs run the SUSE Linux Enterprise High Availability (SLEHA) Extension and the Cray SMW High Availability Extension for SLES 11 SP3 release package, also called the *SMW HA package*.

This document supports the SMW HA SLEHA11SP3.UPO3 software release.

Contents of This Guide

This guide provides the following information:

- [SMW HA Overview](#) on page 7: A summary of the SMW HA system.
- [Install a Cray SMW HA System](#) on page 14: How to perform a full initial installation of the Cray SMW HA system, including the operating system, SLEHA Extension, Cray SMW software, Cray Linux Environment (CLE) software, and the Cray SMW HA software.
- [Customize a Preinstalled SMW HA System](#) on page 204: How to configure a preinstalled SMW HA system with site-specific information such as IP addresses and host names.
- [Upgrade the Cray SMW HA System](#) on page 241: How to upgrade an SMW HA system for a major software release, including the operating system and Cray SMW, CLE, and SMW HA software.
- [Update the Cray SMW HA System](#) on page 325: How to update an SMW HA system for a minor software release, including the Cray SMW, CLE, and SMW HA software.
- [Troubleshooting SMW HA Installation Problems](#) on page 399: How to identify and fix common installation problems with the SMW HA system, including how to disable the SMW HA configuration (in case of problems) and restore a disabled configuration.

For information on managing a running SMW HA system, see *SMW HA XC Administration Guide (S-2551)* .

Related Publications

The following documents contain additional information that may be helpful:

- *SMW HA Release Errata* and the *SMW HA README*, which are provided with the SMW HA release package
- *SLE High Availability Extension SP3 High Availability Guide* from Novell, Inc., which provides information on the SUSE Linux High Availability (SLE HA) Extension software, the Pacemaker Cluster Resource Manager (CRM), and related tools. This document is available online at [suse.com: https://www.suse.com/documentation/sle_ha/](https://www.suse.com/documentation/sle_ha/)

Typographic Conventions

Monospace	Indicates program code, reserved words, library functions, command-line prompts, screen output, file/path names, key strokes (e.g., <code>Enter</code> and <code>Alt-Ctrl-F</code>), and other software constructs.
Monospaced Bold	Indicates commands that must be entered on a command line or in response to an interactive prompt.
<i>Oblique or Italics</i>	Indicates user-supplied values in commands or syntax definitions.
Proportional Bold	Indicates a graphical user interface window or element.
\ (backslash)	At the end of a command line, indicates the Linux® shell line continuation character (lines joined by a backslash are parsed as a single line). Do not type anything after the backslash or the continuation feature will not work correctly.

Scope and Audience

The intended reader of this guide is a system administrator who is familiar with operating systems derived from UNIX.

Feedback

Visit the Cray Publications Portal at <http://pubs.cray.com> and make comments online using the [Contact Us](#) button in the upper-right corner or Email pubs@cray.com. Your comments are important to us and we will respond within 24 hours.

Distribution Media

The Cray SMW SLEHA release includes one DVD or ISO file that contains the Cray SMW HA software package. For an initial installation and most upgrade/update installations, you will also need the release media for the operating system, SMW software, and CLE software.

For more information, see the SMW HA README file provided with the SMW HA release package. Also see the release notes and README files that are provided with the SMW and CLE release packages

SMW HA Overview

The Cray System Management Workstation (SMW) High Availability (HA) system supports SMW failover. An SMW HA system is a Cray XC system with two second-generation high-end SMWs (also called *rack-mount SMWs*) that run the SUSE Linux Enterprise High Availability (SLEHA) Extension and the Cray SMW High Availability Extension (SLEHA) release package. The two SMWs must be installed and configured as specified in this guide.

The SMW failover feature provides improved reliability, availability, and serviceability (RAS) of the SMW, allowing the mainframe to operate correctly and at full speed. This feature adds SMW failover, fencing, health monitoring, and failover notification. Administrators can be notified of SMW software or hardware problems in real time and be able to react by manually shutting down nodes, or allowing the software to manage the problems. In the event of a hardware failure or `rsm` daemon failure, the software will fail over to the passive SMW node, which becomes the active node. The failed node, once repaired, can be returned to the configuration as the passive node.

The SUSE Pacemaker Cluster Resource Manager (CRM) provides administration and monitoring of the SMW HA system with either a command-line interface (`crm`) and a GUI (`crm_gui`). With this interface and associated commands, the SMW administrator can display cluster status, monitor the HSS daemons (configured as cluster resources), configure automatic failover notification by email, and customize the SMW failover thresholds for each resource.

The following topics describe the unique features of the SMW HA system:

- [SMW Cluster Configuration](#) on page 7
- [Shared Storage on the Boot RAID](#) on page 8
- [Storage for the Power Management Database \(PMDB\)](#) on page 9
- [Synchronized Files](#) on page 9
- [Cluster Resources](#) on page 11
- [Limitations of SMW Failover](#) on page 12

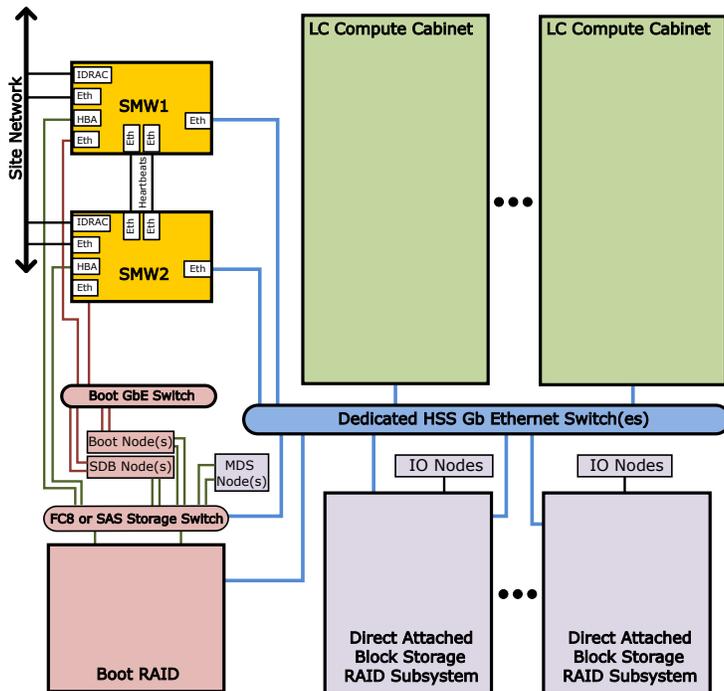
NOTE: The Pacemaker Cluster Resource Manager uses the term *node* to refer to a host in a CRM cluster. On an SMW HA system, a CRM node is an SMW, not a Cray XC compute or service node.

SMW Cluster Configuration

Both SMWs are connected to the boot RAID, and are connected to each other with heartbeat cables between the `eth2` and `eth4` ports on each SMW. The heartbeat connection monitors the health of the cluster. In addition, each SMW is connected to the boot RAID (through FC or SAS cards), to the site network through `eth0`, to the HSS network through `eth1`, and to the boot node through `eth3`. (For more information, see [Network Connections for an SMW HA System](#) on page 16.) An Integrated Dell™ Remote Access Controller (iDRAC) is required on both SMWs.

The following figure shows the major connections between components in an SMW HA system.

Figure 1. SMW HA Hardware Components for a Cray XC System



In a Cray SMW HA cluster, the two SMWs are configured in an active/passive configuration. This configuration lets the passive node take over the SMW functions if a software or hardware fault occurs on the active node. All HSS daemons run on the active SMW. (An additional `stonith` daemon, which monitors SMW health, runs on both SMWs.) At failover, all daemons move to the passive SMW, which then becomes the active one.

During initial installation, the first SMW that is installed and configured becomes the active SMW. The second SMW that is installed and configured becomes the passive SMW. However, either SMW can be active during normal operation. The cluster configuration does not remember which SMW was initially configured to be active.

Shared Storage on the Boot RAID

The SMW HA system uses shared disk devices on the boot RAID for data that must be highly available. The shared directories are mounted only on the active SMW. When a failover occurs, access to these directories is automatically transferred to the other SMW as part of the failover process.

IMPORTANT: Because several file systems are shared between the two SMWs, an SMW HA system has a slightly increased risk for double-mount problems. Do **not** mount the CLE boot root, the shared root, or any other CLE file systems from the boot RAID on both SMWs at the same time.

The SMW HA system uses shared space on the boot RAID for the following directories:

`/var/opt/cray/disk/1` Log disk. The following directories symbolically link to the Log disk:

- `/var/opt/cray/debug`
- `/var/opt/cray/dump`
- `/var/opt/cray/log`

<code>/var/lib/mysql</code>	MySQL HSS database. Although the database is shared, the HSS database server runs on the active SMW only.
<code>/home</code>	SMW home directories.
<code>/var/lib/pgsql</code>	Power Management Database (PMDB), if on the shared boot RAID. Note that mirrored storage is preferred. For more information, see Storage for the Power Management Database (PMDB) on page 9.

Storage for the Power Management Database (PMDB)

The Power Management Database (PMDB) is a PostgreSQL database that contains power management data, event router file system (erfs) data, and optional System Environment Data Collections (SEDC) data. The directory `/var/lib/pgsql` is the mount point for the PMDB. On an SMW HA system, this directory should be configured to be available after failover. When a failover occurs, access to `/var/lib/pgsql` is automatically transferred to the other SMW as part of the failover process.

The following options are available for PMDB storage:

- **Mirrored storage (preferred):** An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (eth5) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.

For more information, see [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172.

- **Shared storage:** A logical disk, configured as a LUN (Logical Unit) or logical volume on the boot RAID. The boot RAID must have sufficient space for `/var/lib/pgsql`.

For more information, see [Shared Storage for SMW HA](#) on page 15 and [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177.

- **Unshared storage (not recommended):** Each SMW stores an unsynchronized copy of `/var/lib/pgsql` on the local root disk. Cray strongly recommends using either mirrored storage (preferred) or shared storage. An unshared PMDB is split across both SMWs; data collected before an SMW failover will be lost or not easily accessible after failover.

Synchronized Files

For files not located on the shared storage device, the SLEHA Extension software includes the `csync2` utility to synchronize (*sync*) important files between the two SMWs. When a file changes on the active SMW, it is automatically synchronized to the passive SMW.

File synchronization is automatically configured during initial installation. The file `/etc/csync2/csync2_cray.cfg` lists the Cray-specific files and directories that must be synchronized, as well as small files that are convenient to keep in sync.

File synchronization happens in one direction only: from the active SMW to the passive SMW. If you change a synchronized file on the passive SMW, the change will not be propagated to the active SMW in the course of

normal operations and could be overwritten on the passive SMW later if there is a subsequent change to the corresponding file on the active SMW. However, if a failover occurs, the previously passive SMW becomes the active SMW. If the change is still in place, the changed file becomes a candidate for propagation to the other SMW (subject to the rules of file conflict resolution).

The `fsync` resource controls file synchronized operations. Every 100 seconds, `fsync` checks for files that need to be synchronized.

IMPORTANT: If a failover occurs before a file synchronization operation has completed, it could result in the loss of the latest updates.

The `csync2` utility synchronizes the required files and directories for the SMW HA cluster, such as `/etc/passwd` and `/opt/cray/hss/*/etc/*`. For more information, see [About File Synchronization Between HA SMWs](#) or examine the contents of `/etc/csync2/csync2_cray.cfg`.

Very large files are explicitly excluded from synchronization (such as `/opt/cray/hss-images/master`). The `csync2` utility is designed to synchronize small amounts of data. If `csync2` must monitor many directories or synchronize a large amount of data, it can become overloaded and failures may not be readily apparent. Cray recommends that you do not change the list of synchronized files (or add only small files); copy large files and directories manually to the other SMW. For more information, see [About File Synchronization Between HA SMWs](#).

Add Site-specific Files to the Synchronization List

The file `/etc/csync2/csync2_cray.cfg` specifies the Cray-specific files and directories that must be synchronized, as well as small files that are convenient to keep in sync.

IMPORTANT: The `csync2` utility is designed to synchronize small amounts of data. If `csync2` must monitor many directories or synchronize a large amount of data, it can become overloaded and failures may not be readily apparent. Cray recommends that you add only small files to `/etc/csync2/csync2_cray.cfg`. For example, do not synchronize the following files or directories:

- `/home`
- `/home/crayadm/.ssh/authorized_keys`
- `/opt/xt-images` (Cray boot images are very large)
- `/tmp/SEDC_FILES`, if SEDC does not use the PMDB
- `/etc/hosts`
- Very large files

TIP: You can use `scp` to copy a large, static file to the passive SMW, as in this example:

```
smw1:~ # scp -pr /path/file smw2:/path/file
```

For directories and files that may change during the copy operation, you can use the `rsync` command.

1. For each file or directory on the active SMW that you want to synchronize, ensure that the parent directory exists on the passive SMW. In some cases, you must either manually create directories on the passive SMW or copy the directory structure from the active SMW. With either method, be sure that owner, group, and permissions are maintained, because `csync2` can be sensitive to mismatches.
2. Edit the file `/etc/csync2/csync2_cray.cfg` as `root` on the active SMW.

- To add a file or directory, add the full path (one entry per line) to `/etc/csync2/csync2_cray.cfg`. Comments in this file explain how to make changes.

IMPORTANT: For a symbolic link, only the link itself is synchronized, not the content (destination) of the symbolic link.

- Save your changes and exit the editor.

The `fsync` resource will synchronize the additional files and directories the next time it runs.

- If there are local changes to `/etc/hosts` on `smw1`, manually copy `/etc/hosts` to `/etc/hosts` on `smw2`. The customized entries must be above the first section of "XT Cabinet x - y".

```
smw2:~ # cp /etc/hosts /etc/hosts.sav
smw2:~ # scp smw1:/etc/hosts /etc/hosts
```

Then edit the `/etc/hosts` file on `smw2`:

- Change IP addresses `10.1.1.x`, `10.2.1.x`, `10.3.1.x`, and `10.4.1.x` to `10.1.1.y`, `10.2.1.y`, `10.3.1.y`, and `10.4.1.y` where if `x` is 2 `y` is 3 and if `x` is 3 `y` is 2.
- Change the line `smw1-ip smw1 smw1` to `smw2-ip smw2 smw2`.

Cluster Resources

A resource is any type of service or application that is managed by the Pacemaker Cluster Resource Manager, such as a daemon or file system. In an SMW HA system, the HSS (rsms) daemons are configured as resources.

Each time a resource fails, it is automatically restarted and its failcount is raised. If the failcount exceeds the defined migration threshold for the resource, a failover occurs and management of all cluster resources migrates to the other SMW, making it the active SMW. The original SMW will no longer be allowed to run the failed resource, so no failback can occur until the resource's failcount is reset for that SMW.

TIP: You can reset failcounts with the `clean_resources` or `clear_failcounts` command. For more information, see [Resources Are Stopped](#).

An SMW HA system includes the following resources:

<code>ClusterIP</code> , <code>ClusterIP1</code> , <code>ClusterIP2</code> , <code>ClusterIP3</code> , and <code>ClusterIP4</code>	Control and monitor the Ethernet connections (<code>eth0</code> , <code>eth1</code> , <code>eth2</code> , <code>eth3</code> , and <code>eth4</code> , respectively).
<code>ClusterMonitor</code>	Records failcounts and failed actions in the log file <code>/var/log/smwha.log</code> at cluster startup, then clears the failure data from <code>crm</code> (for example, in the output of <code>crm_mon -r1</code>).
<code>ClusterTimeSync</code>	Monitors the kernel time on each SMW. If the difference is greater than 60 seconds, both SMWs are synchronized with the NTP server. If the time difference is greater than 10 hours, the time is not synchronized.
<code>cray-syslog</code>	Controls and monitors Lightweight Log Management (LLM).

dhcpcd	Controls and monitors <code>dhcpcd</code> as used by the SMW HA feature.
fsync	Provides file synchronization using <code>csync2</code> .
homedir	Mounts and unmounts the shared <code>/home</code> directory.
hss-daemons	Controls and monitors HSS daemons; corresponds to the <code>/etc/init.d/rsms</code> startup script.
ip_drbd_pgsql	Controls and monitors the Ethernet connection (<code>eth5</code>) between the two SMWs for Power Management Database (PMDB) mirrored storage using a Distributed Replicated Block Device (DRBD).
ms_drbd_pgsql	Monitors the master/slave DRBD cluster resource for PMDB mirrored storage.
Notification	Provides automatic notification email when a failover occurs.
md-fs	Mounts, unmounts, and monitors the shared MySQL database, <code>/var/lib/mysql</code> .
ml-fs	Mounts, unmounts, and monitors the shared log directory, <code>/var/opt/cray/disk/1</code> , which symbolically links to the <code>dump</code> , <code>install</code> , and <code>log</code> subdirectories in <code>/var/opt/cray/</code> .
mysqld	Controls and monitors MySQL.
pm-fs	Controls and monitors the Power Management Database (PMDB) file system, <code>/var/lib/pgsql</code> .
postgresqld	Controls and monitors the Power Management Database (PMDB) PostgreSQL server, <code>postgresqld</code> .
stonith-1 and stonith-2	Monitors the health of the other SMW. Each SMW monitors its peer and has the ability to power off that peer at failover time using the STONITH capability. STONITH failovers are used when the state of the failing SMW cannot be determined. A STONITH failover powers off the failing SMW to guarantee that the newly active SMW has exclusive access to all cluster managed resources.

Limitations of SMW Failover

The SMW HA failover feature has the following limitations:

- Both SMWs must be of the same type, either Dell 815s or Dell 630s.
- Both SMWs must run the same versions of SLES and SMW/HSS software.
- System administration of an SMW HA environment is more complex than administration of a system with a single SMW.

-
- Before using a command that interacts with the HSS daemons, wait for 30 - 60 seconds after failover to ensure that all cluster resources have started. In the first 30 seconds after failover, resources may appear to be started, then change to another state. Although one might be able to log in via the virtual IP address before this period is over, the cluster is not ready for use until all resources are fully started.

TIP: Use `crm_mon` to verify that all cluster resources have started after failover. For more information, see [crm_mon](#).

- SMW and CLE upgrades in an HA environment require some duplication of effort, with portions of the procedure done individually to each SMW. System down-time requirements for operating system upgrades are somewhat longer as a result.
- There is no support for seamless failover (also called *double failure*) if errors occur while the system is doing error handling for another system component. If an HSS daemon or other SMW process were doing some type of error handling that got interrupted by an (unrelated) failover, when that daemon restarts on the new SMW it may not be able to resume operation where it left off and complete the recovery from the first error. In this case, even though a failover occurs, manual intervention might still be required to return the system to an operational state.
- There is no support for seamless failover during operational commands. All user commands that were started from the active SMW are terminated. These commands must be restarted on the new active SMW. The restarted commands might not start with the same internal states, if those commands do not provide persistent capabilities. An interrupted operation such as `xtbootsys`, `shutdown`, `dump`, `warm-swap`, or `flash` will need to be reissued after failover has completed and the other SMW becomes active.
- Partial migration of managed resources is not supported. For example, the SMW HA system does not support migration of individual HSS daemons or resources to the other SMW. A particular SMW is either *active*, with complete responsibility for all HSS daemons, or *passive* with no HSS daemons running.
- If both SMWs are started (powered on) at the same time, a race condition can develop that could result in one SMW being powered off via the STONITH capability. Before starting the second SMW, wait until the first SMW has completed startup and initialized all cluster resources.
- During failover, if there is no communication between the SMW and the Cray mainframe for about 30 seconds, workload throttling can occur; therefore, auto-throttling of applications is likely while an actual SMW failover is taking place. Blades begin to auto-throttle if essential HSS daemons (`erd`, `state-manager`, or `xtnlrd`) are unavailable and lasts until those daemons resume operation on the other SMW. On a single-cabinet system, the throttled period was fairly consistent, lasting 37 seconds. The throttled period may increase for larger systems.
- Direct Attached Lustre (DAL) is not supported with the SMW HA failover release.
- If the Power Management Database (PMDB) is on local SMW disks rather than on mirrored or shared storage, PMDB data collected before an SMW failover will be lost or not easily accessible after failover.

Install a Cray SMW HA System

IMPORTANT:

For a new SMW HA system: Cray ships systems with installation and most of the configuration completed. Unless you need to reinstall the SMW HA system, skip these initial installation procedures. Instead, use the procedures in [Customize a Preinstalled SMW HA System](#) on page 204.

A complete initial installation for a new Cray SMW HA system includes installing the operating system and the Cray SMW, CLE, and SMW HA software on both SMWs. The complete installation includes the following tasks:

1. [Prepare to Install a New SMW HA System](#) on page 14: Plan shared storage, identify network connections, gather site-specific configuration values, and note the default passwords.
2. [Install SMW Software on the First SMW](#) on page 18
3. [Configure the Boot RAID for SMW HA](#) on page 47
4. [Install CLE Software on the First SMW](#) on page 65
5. [Install SMW Software on the Second SMW](#) on page 141
6. [Install CLE Software on the Second SMW](#) on page 157
7. [Install SMW HA Software](#) on page 158
8. [Configure the Cluster](#) on page 161
9. [Back Up a Newly-installed SMW HA System](#) on page 182
10. [Change Default HA Passwords After Installation](#) on page 201

This guide uses the following conventions to refer to these SMWs:

- The host name `smw1` specifies the SMW that is configured to be the first active SMW during initial installation. In examples, the prompt `smw1:~ #` shows a command that runs on this SMW.
- The host name `smw2` specifies the SMW that has been configured to be the first passive SMW during initial installation. In examples, the prompt `smw2:~ #` shows a command that run on this SMW.
- The virtual host name `virtual-smw` specifies the currently active SMW (which could be either `smw1` or `smw2`). This virtual host name is defined during installation and configuration.

Prepare to Install a New SMW HA System

- Read the *SMW HA Release Notes*, the *SMW HA README*, and the *SMW HA Release Errata* to confirm the required versions for the operating system, SMW, and SMW HA software and to determine if there are any additional installation-related requirements, corrections to these installation procedures, and other relevant information about the release package.
- Read the Field Notices (FNs) to identify whether there are any changes to this release package or the installation instructions.

- Read any Field Notices (FNs) related to kernel security fixes.

IMPORTANT: Kernel 3.0.101-0.461 (provided in FN-6029) or later is required for an SMW HA system. There is a SLES kernel dependency on the `ocfs2-kmp-default` RPM package that will prevent some SLES HA RPMs from being installed unless this kernel update has been applied.

- Read this section before you start the installation to ensure the following:
 - Ensure that the prerequisites are satisfied before beginning an initial SMW software installation.
 - Verify that the two SMWs are correctly cabled and that network connections are in place (see [Network Connections for an SMW HA System](#) on page 16).
 - Identify the configuration values for the system (see [Configuration Values for an SMW HA System](#) on page 16).
- Plan space on the boot RAID for the shared storage for the SMW HA system before installing the SMW and CLE software. For more information, see [Shared Storage on the Boot RAID](#) on page 8.
- For an existing system: Back up the current SMW software before installing the SMW and SMW HA packages.

NOTE: Cray recommends removing old SMW log files to reduce the amount of time needed to back up the SMW.

Shared Storage for SMW HA

An SMW HA system requires boot RAID LUNs (Logical Units) for shared storage for the cluster, in addition to the LUNs required for a Cray system with a single SMW. Before installing the SMW and CLE software, plan space on the boot RAID for the shared directories.

Refer to [Recommended Boot RAID LUN Values](#) on page 68 for the basic set of required LUNs.

The following table shows the minimum partition sizes for the additional LUNs for an SMW HA system. A large system may require additional space for the shared directories; review the requirements of the system in order to determine the appropriate size for these LUNs.

Table 1. Minimum Boot RAID LUN Sizes for SMW Failover

Directory on SMW	Description	Minimum Size
<code>/var/lib/mysql</code>	MySQL HSS database	150GB
<code>/var/opt/cray/disk/1</code>	Log disk. The following directories symbolically link to the Log disk: <ul style="list-style-type: none"> <code>/var/opt/cray/debug</code> <code>/var/opt/cray/dump</code> <code>/var/opt/cray/log</code> 	500GB
<code>/home</code>	SMW home directories	500GB
<code>/var/lib/pgsql</code>	Power Management Database (PMDB), if on the shared boot RAID	500GB

If the PMDB will use shared storage on the boot RAID, check the size of `/var/lib/pgsql` and make sure that RAID disk has enough space to hold the PMDB data. Use the following command to display the size of the PMDB:

```
smw1:~ # du -hs /var/lib/pgsql
```

IMPORTANT: Cray strongly recommends using mirrored storage, if available, for the PMDB. For more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9.

Network Connections for an SMW HA System

An SMW HA system uses `eth2` and `eth4` (on the second Ethernet card) for heartbeat connections to the other SMW, in addition to the network connections required for single SMW.

Each SMW must be connected to the customer network through `eth0`, to the HSS network through `eth1`, to the boot node through `eth3`, and to the boot RAID through the Fibre Channel (FC) or SAS card. In addition, `eth2` and `eth4` must directly connect the two SMWs to each other as heartbeat monitoring channels.

Each SMW must have the following private network connections:

- `eth0` - To the customer network
- `eth1` - To the HSS network
- `eth2` - To the other SMW (heartbeat connection)
- `eth3` - To the boot node
- `eth4` - To the other SMW (redundant heartbeat connection)
- `eth5` - To the mirrored storage for the power management database (PMDB), if available (see [Storage for the Power Management Database \(PMDB\)](#) on page 9)

For more information on the network connections for an SMW, see [Network Connections](#) on page 22.

Configuration Values for an SMW HA System

An SMW HA cluster uses the following fixed IP addresses. These IP addresses are set by default and are not site dependent.

Table 2. Fixed IP Addresses for an SMW HA system

IP Address	Description
10.1.0.1	Primary boot RAID controller
10.1.0.2	Secondary boot RAID controller
10.1.0.15	Storage RAID controller
10.1.1.1	SMW, <code>eth1</code> - Virtual <code>eth1</code> connection
10.1.1.2	SMW, <code>eth1</code> - Actual <code>eth1</code> connection for <code>smw1</code>
10.1.1.3	SMW, <code>eth1</code> - Actual <code>eth1</code> connection for <code>smw2</code>
10.2.1.1	SMW, <code>eth2</code> - Virtual primary heartbeat connection for SMW failover
10.2.1.2	SMW, <code>eth2</code> - Actual primary heartbeat connection for <code>smw1</code>
10.2.1.3	SMW, <code>eth2</code> - Actual primary heartbeat connection for <code>smw2</code>
10.2.1.0	Network address to bind to (for <code>eth2</code> primary heartbeat connection on <code>smw2</code>)

IP Address	Description
10.3.1.1	SMW, eth3 - Virtual eth3 connection
10.3.1.2	SMW, eth3 - Actual eth3 connection for smw1
10.3.1.3	SMW, eth3 - Actual eth3 connection for smw2
10.4.1.1	SMW, eth4 - Virtual redundant heartbeat connection for SMW failover
10.4.1.2	SMW, eth4 - Actual redundant heartbeat connection for smw1
10.4.1.3	SMW, eth4 - Actual redundant heartbeat connection for smw2
10.4.1.0	Network address to bind to (for eth4 primary heartbeat connection on smw2)
10.5.1.2	SMW, eth5 - Mirrored PMDB disk connection for smw1
10.5.1.3	SMW, eth5 - Mirrored PMDB disk connection for smw2
127.0.0.1	Localhost (loopback)
225.0.0.1	Multicast IP address for eth4
226.0.0.1	Multicast IP address for eth2
1694	Multicast port for primary heartbeat connection (for eth2 and eth4 on smw2)

An SMW HA system also requires the following site-dependent host names and IP addresses. You may find it helpful to record the actual values for the site.

NOTE: This table lists the HA-specific values only. For the other site-dependent values that apply to all systems (with either one or two SMWs), see [Configuration Values](#) on page 22.

Table 3. Site-dependent Configuration Values for an SMW HA System

Description	Example	Actual Value
Virtual host name for SMW HA cluster	virtual-smw	
Host name for first SMW	smw1	
Host name for second SMW	smw2	
iDRAC host name on first SMW	smw1-drac	
iDRAC host name on second SMW	smw2-drac	
Customer network IP address for virtual SMW (the SMW HA cluster)	173.31.73.165	
IP address for first SMW	173.31.73.60	
IP address for second SMW	173.31.73.61	
iDRAC IP address on first SMW	172.31.73.77	

Description	Example	Actual Value
iDRAC IP address on second SMW	172.31.73.79	

IMPORTANT: The IP addresses for the virtual SMW HA cluster (`virtual-smw`) and the actual SMWs `smw1` and `smw2`) must be on the same subnet.

Passwords For an SMW HA System

On an SMW HA system, the `root` password must be the same on each SMW. In addition, the Integrated Dell™ Remote Access Controllers (iDRAC) password on each SMW must be the same as the `root` password. If you change the default `root` password after installing the SMW software, ensure that the iDRAC `root` accounts use the same password.

The SMW HA configuration process uses the `root` password for the several cluster management accounts (the `hacluster` user, the `stonith-1` resource, and the `stonith-2` resource). When changing passwords after installing the SMW HA software, ensure that these accounts also use the same password as `root` on both SMWs.

[Default Passwords for an SMW HA System](#) lists the default values for the passwords that must be the same on both SMWs.

Table 4. Default Passwords for an SMW HA System

ID	Default Password
<code>root on smw1</code>	<code>initial0</code>
<code>root on smw2</code>	<code>initial0</code>
<code>root (iDRAC) on smw1</code>	<code>initial0</code>
<code>root (iDRAC) on smw2</code>	<code>initial0</code>
<code>hacluster (for logging in to crm_gui)</code>	same as <code>root</code> password (set during HA configuration)
<code>stonith-1 resource</code>	same as <code>root</code> password (set during HA configuration)
<code>stonith-2 resource</code>	same as <code>root</code> password (set during HA configuration)

For more information, see [Change Passwords on an SMW HA System](#).

Install SMW Software on the First SMW

For the first SMW in the SMW HA cluster, the procedures to install the Cray SMW software are almost the same as for a single SMW, with a few HA-specific differences.

Installing the SMW software on the first SMW includes the following tasks:

1. [Prepare to Install SMW Software on the First SMW](#) on page 19.

-
2. [Install the SMW Release Package](#) on page 19: Install the base operating system and Cray SMW software, perform hardware discovery and power-up, confirm SMW communication, change default passwords, set up SUSE firewall and IP tables, and configure the Simple Event Coordinator (SEC).

Note that for the first SMW, the procedure to back up the newly-installed SMW software can wait until after both SMWs have been installed and the SMW HA configuration is complete.

3. [Finish Installing SMW Software on the First SMW](#) on page 46.

IMPORTANT: If you are converting an existing Cray system (with a single SMW) to an SMW HA cluster, you do not need to reinstall the operating system and the full SMW software. Instead, upgrade or update the existing SMW to the required SMW and CLE release software (see [Upgrade the Cray SMW HA System](#) on page 241 or [Update the Cray SMW HA System](#) on page 325). Then continue with these procedures:

- [Configure the Boot RAID for SMW HA](#) on page 47
- [Install SMW HA Software](#) on page 158
- [Configure the Cluster](#) on page 161

Prepare to Install SMW Software on the First SMW

Before installing the SMW software, use this procedure to prepare the first SMW.

1. Choose the first SMW. You can begin the installation on either SMW. The SMW that is installed first will initially become the active SMW when the SMW HA cluster is fully configured. The examples in the installation procedures show the host name `smw1` for the first SMW.
2. Verify the heartbeat connections between the two SMWs. Two Ethernet ports are used for heartbeat connections between the two SMWs: `eth2` (on the first quad Ethernet card) and `eth4` (on the second quad Ethernet card), as described in [Network Connections for an SMW HA System](#) on page 16.
3. Ensure that the boot RAID is disconnected.

When installing the operating system, only the boot disk should be connected to the SMW. All other internal disks should be disconnected (ejected). The boot RAID **must** be disconnected to prevent data corruption when installing the operating system.

Install the SMW Release Package

Follow these procedures to perform an initial or clean installation of the Cray System Management Workstation (SMW) 7.2.UP04 release package.

Cray provides two rack-mount SMW models: the Dell PowerEdge™ R815 Rack Server and the Dell PowerEdge™ R630 Rack Server. The figure below shows an easy way to distinguish between the two rack-mount models when viewing them from the front.

Figure 2. Distinguishing Features of Dell R815 and R630 Servers



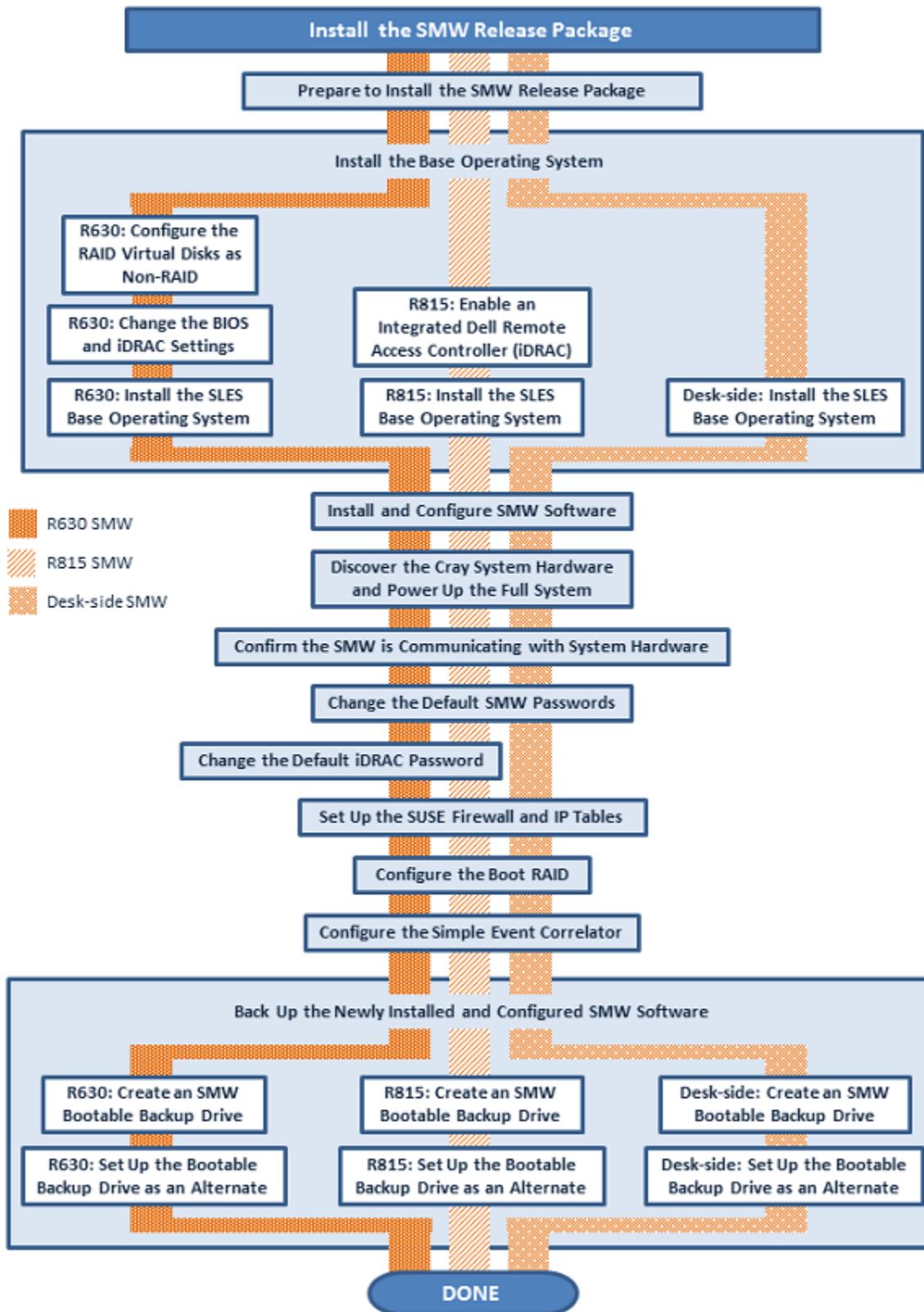
Dell R815: 2U high and 6 drive bays



Dell R630: 1U high and 8 drive bays

Most of the installation procedures that follow apply to all SMW models. Those that apply to only one or more of the models are clearly marked, and the visual guide below also indicates applicability.

Figure 3. Visual Guide to SMW Release Package Installation



Prepare to Install the SMW Release Package

- Read the *SMW Release Errata* and the *SMW README* provided with the SMW release package for any additional installation-related requirements, corrections to this installation guide, and other relevant information about the release package.
- Read the Field Notices (FN) related to kernel security fixes to identify any changes to this release package. Apply any needed changes before installing the new software.
- Use the correct publication. To install the SMW release package on a system configured for SMW high availability (HA) with the SMW failover feature, use *SMW HA XC Installation Guide* (S-0044). Otherwise, use *System Management Workstation (SMW) Software Installation Guide* (S-2480).
- Verify that the network connections are in place (see [Network Connections](#) on page 22).
- Know which configuration values are site-specific and which are defaults (see [Configuration Values](#) on page 22).
- Refer to the default passwords used during the installation process (see [Passwords](#) on page 24).
- Become familiar with the mapping of Linux device names to physical drive slots (see [Rack-mount SMW: Linux /dev Names Mapped to Physical Drive Slots](#)).

Network Connections

The following network connections are required.

- A standalone SMW with a single quad-ethernet card has these private network connections:
 - eth0 - To the customer network
 - eth1 - To the Hardware Supervisory System (HSS) network
 - eth2 - Reserved for SMW failover
 - eth3 - To the boot node
- An SMW configured for SMW failover has a second quad-ethernet card with these connections:
 - eth4 - Used for SMW failover
 - eth5 - Used for mirrored storage for the power management database (PMDb), if available
 - eth6-7 - Reserved for future use

NOTE: Ethernet port assignments are valid only after the SMW software installation completes (see [R815 SMW: Install the SLES Base Operating System](#) on page 24).

- The SMW must have a Fibre Channel or Serial Attached SCSI (SAS) connection to the boot RAID.
- The boot node must have a Fibre Channel or SAS connection to the boot RAID.
- The service database (SDB) node must have a Fibre Channel or SAS connection to the boot RAID.

IMPORTANT: The SMW must be disconnected from the boot RAID before the initial installation of the SLES software. Ensure that the Fibre Channel optic cable connectors or SAS cable connectors have protective covers.

Configuration Values

The following IP addresses are set by default and are not site dependent.

NOTE: These default IP addresses are only for a standalone SMW. For an SMW HA system, see the default IP addresses in [Configuration Values for an SMW HA System](#) on page 16.

Table 5. Default IP Addresses

IP Address	Description
10.1.0.1	Primary boot RAID controller
10.1.0.2	Secondary boot RAID controller
10.1.0.15	Storage RAID controller
10.1.1.1	SMW, eth1
10.2.1.1	SMW, eth2 - Reserved for SMW failover
10.3.1.1	SMW, eth3
10.3.1.254	boot node
10.4.1.1	SMW, eth4 - Reserved for SMW failover
127.0.0.1	localhost (loopback)

The following configuration values are site dependent. Record the actual values for the installation site in the third column. References to rack-mount SMW include both the Dell R815 and Dell R630 models..

NOTE: In addition to these values, there are HA-specific values that apply to an SMW HA system. See [Configuration Values for an SMW HA System](#) on page 16.

Table 6. Site-dependent Configuration Values

Description	Example	Actual Value
SMW hostname	xtsmw	
Domain	cray.com	
Aliases	cray-smw smw01	
Customer network IP address	192.168.78.68	
Customer network netmask	255.255.255.0	
Default gateway	192.168.78.1	
Domain names to search	us.cray.com mw.cray.com	
Nameserver IP address	10.0.73.30 10.0.17.16	
For rack-mount SMW only: iDRAC hostname	cray-drac	
For rack-mount SMW only: iDRAC IP address	192.168.78.69	
For rack-mount SMW only: iDRAC Subnet Mask	255.255.255.0	
For rack-mount SMW only: iDRAC Default GW	192.168.78.1	
Timezone	US/Central	
NTP servers	ntpghost1 ntpghost2	

Description	Example	Actual Value
X dimension	1-64	
Y dimension	Cray XE and Cray XK systems: 1-16; Cray XC Series Systems: 1-32	
Topology Class	Cray XE and Cray XK systems: 0, 1, 2, 3; Cray XC Series Systems: 0, 2 NOTE: Regardless of the number of cabinets in the system, Cray XC Series air-cooled systems must be set to 0. Cray XC Series liquid-cooled systems can be class 0 or 2.	

Passwords

The following default account names and passwords are used throughout the SMW software installation process. Cray recommends changing these default passwords after completing the installation.

Table 7. Default System Passwords

Account Name	Password
root	initial0
crayadm	crayadm
cray-vnc	cray-vnc
mysql	None; a password must be created
admin (DDN™ boot RAID)	password
user (DDN boot RAID)	password
admin (DDN storage RAID)	password
user (DDN storage RAID)	password
root (iDRAC)	initial0

Back Up the Current Software

Before installing the new SMW software, back up the current SMW software installation.

- For a Dell R630 (rack-mount) SMW, use [R630 SMW: Create an SMW Bootable Backup Drive](#) on page 192.
- For a Dell R815 (rack-mount) SMW, use [R815 SMW: Create an SMW Bootable Backup Drive](#) on page 182.

R815 SMW: Install the SLES Base Operating System

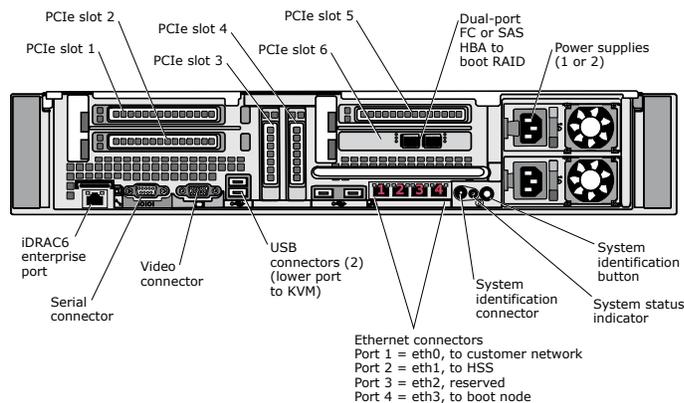
This procedure describes how to install SLES11SP3 as the base operating system on an R815 SMW. Use the DVD labeled `Cray-SMWbase11SP3-` to perform this installation.

1. If the SMW is up, `su` to `root` and shut it down.

```
crayadm@smw> su - root
smw# shutdown -h now;exit
```

2. Disconnect the SMW connection to the boot RAID; disconnect the data cables and place protective covers on the Fibre Channel or SAS cables and connectors (if present).

Figure 4. Dell R815 SMW Rear Connections



3. Eject all the disk drives except for the primary disk in slot 0.
4. Power up the SMW. When the BIOS power-on self-test (POST) process begins, quickly press the F2 key after the following messages appear in the upper-right of the screen.

```
F2 = System Setup
F10 = System Services
F11 = BIOS Boot Manager
F12 = PXE Boot
```

When the **F2** keypress is recognized, the `F2 = System Setup` line changes to `Entering System Setup`.

After the POST process completes and all disk and network controllers have been initialized, the BIOS set-up screen appears.

5. Use the down-arrow key to highlight **Boot Settings**. Press the **Enter** key.

A window listing the following appears:

```
Boot Mode ..... BIOS
Boot Sequence ..... <ENTER>
USB Flash Drive Emulation Type..... <ENTER>
Boot Sequence Retry ..... <Disabled>
```

6. Use the down-arrow key to highlight **Boot Sequence**. Press the **Enter** key.

A window listing the following appears.

```

√ 1. Hard Drive C: (Integrated SAS 500 ID0A LUN0 ATA)
√ 2. Virtual CD
√ 3. Sata Optical Drive
√ 4. Embedded NIC 1 MBA v6.0.11 Slot 0100
√ 5. Virtual Floppy

```

7. Using the up-arrow and down-arrow keys to select, and using the space key to enable/disable entries, modify the list so that only the 1. `Hard Drive C:` entry has a check mark.

```

√ 1. Hard Drive C: (Integrated SAS 500 ID0A LUN0 ATA)
  2. Virtual CD
  3. Sata Optical Drive
  4. Embedded NIC 1 MBA v6.0.11 Slot 0100
  5. Virtual Floppy

```

8. Press the **Esc** key to exit the **Boot Sequence** window.
9. Press the **Esc** key again to exit the **Boot Settings** window.
10. Insert the base operating system DVD labeled `Cray-SMWbase11SP3-` into the CD/DVD drive. (The DVD drive on the front of the SMW may be hidden by a removable decorative bezel.)
11. Press the **Esc** key a final time to save changes and exit the BIOS set up. A screen listing exit options appears.

```

Save changes and exit
Discard changes and exit
Return to Setup

```

12. Ensure that `Save changes and exit` is highlighted. Then press the **Enter** key. The SMW resets automatically.
13. When the BIOS POST process begins again, within 5 seconds, press the **F11** key after the following messages appear in the upper-right of the screen.

```

          F2 = System Setup
          F10 = System Services
          F11 = BIOS Boot Manager
          F12 = PXE Boot

```

When the **F11** keypress is recognized, the `F11 = BIOS Boot Manager` line changes to `Entering BIOS Boot Manager`.

14. Watch the screen carefully as text scrolls until `iDRAC6 Configuration Utility 1.57` appears. Within 5 seconds, press the **Ctrl-E** key when the prompt `Press <Ctrl-E> for Remote Access Setup within 5 sec...` displays.

```

  0   5   0   ATA   WDC WD5000BPVT-0 1A01   465 GB
LSI Corporation MPT2 boot ROM successfully installed!

iDRAC6 Configuration Utility 1.57
Copyright 2010 Dell Inc. All Rights Reserved

iDRAC6 Firmware Revision version: 1.54.15
Primary Backplane Firmware Revision 1.07

```

```

-----
                IPv6 Settings
-----

IPv6 Stack      : Disabled
Address 1      : ::
Default Gateway : ::
-----

                IPv4 Settings
-----

IPv4 Stack      : Enabled
IP Address     : 172. 31. 73.142
Subnet mask    : 255.255.255. 0
Default Gateway : 172. 31. 73. 1
Press <Ctrl-E> for Remote Access Setup within 5 sec...

```

The **iDRAC6 Configuration Utility** window appears.

15. In the **iDRAC6 Configuration Utility**, select **Virtual Media Configuration**.

The **Virtual Media Configuration** window appears.

16. Select the **Virtual Media** line until it indicates **Detached**.

```

Virtual Media ..... Detached

```

17. Press the **Esc** key to exit the **Virtual Media Configuration** window.

18. Press the **Esc** key again to exit the **iDRAC6 Configuration Utility** window.

The **BIOS Boot Manager** screen appears.

19. Use the up-arrow and down-arrow keys to highlight the **SATA Optical Drive** entry.

20. Press the **Enter** key to boot from the installation DVD.

21. Within 10 to 15 seconds after the SUSE Linux Enterprise Server boot menu displays, use the down-arrow key to scroll down and select the **Cray SMW Initial Install Rackmount** option.

```

- Boot from Hard Disk
- Cray SMW Initial Install
- Cray SMW Initial Install Rackmount
- Cray SMW Upgrade Install
- Repair Installed System
- Rescue System
- Check Installation Media

```

Then press the **Enter** key.

If the timeout ends before a selection is made, the system will boot from the hard disk (the default selection). If that happens, shut down the SMW, then begin the power-up sequence again.

As the base installation progresses, the following phrases appear on the screen:

```

Analyzing Computer
System Probing

```

Preparing System for Automated Installation Installation Settings

After these screens are displayed, the installation pauses on `Installation Settings`.

22. Partition the drive for installation of the base operating system and SMW software.

- a. Remove and recreate the `sda` partitions.
 1. For each partition of `sda`, double-click on the `Device name` of that specific `Hard Disk` table entry, which opens the `Partition: /dev/sd...` table for that drive and which has `Edit`, `Resize` and `Delete` buttons below the table.
 2. Select the **Delete** button. A pop-up window indicating `Really delete /dev/sd...` appears.
 3. Verify the device, and select **Yes** to delete the existing disk partition on that drive. The specific `/dev/sd...` entry will be removed from the displayed list.
 4. Repeat these steps for all entries associated with the `sda` disk until the displayed list is empty.
- b. Create new partitions for `swap` and `root` for the `sda` device sized for this system. This is the target boot device for the operating system installation.
 1. Select the **Add** button at the bottom of the **Hard Disk: /dev/sd...** window, which opens the **Add Partition on /dev/sd...** screen.
 2. Select **Primary Partition**, then select **Next**. The **New Partition Size** screen displays.
 3. Select **Custom** size, then enter the swap partition size into the box. This swap partition size should be the same as the total system RAM memory on the SMW.

Enter either **8GB** or **32GB**, depending on the total system RAM memory on the SMW.

TIP: For an SMW that has one power supply, enter **8GB**. For an SMW that has two power supplies, enter **32GB**.

Then select **Next**. The **Formatting Options and Mounting Options** screen displays.

4. Select **Format partition** and select **Swap** from the pull-down menu under **File system**.
5. Select **Mount partition** and select **Swap** from the pull-down menu under **Mount Point**.
6. Select the **Finish** button. The **Hard Disk: /dev/sd...** window displays again, but with the new partition entry.
7. Select the **Add** button at the bottom of the window, which opens the **Add Partition on /dev/sd...** screen.
8. Select **Primary Partition**, then select **Next**. The **New Partition Size** screen displays.
9. Select **Maximum Size**, then select **Next**. The **Formatting Options and Mounting Options** screen displays.
10. Select **Format partition** and select **Ext3** from the pull-down menu under **File system**.
11. Select **Mount partition** and select `/` from the pull-down menu under **Mount Point**.
12. Select **Finish**. Partitioning of the target boot device for the operating system installation has been completed. On the **Expert Partitioner** screen, select **Accept** to accept the changes. A pop-window stating the following will be displayed:

**Changes in disk partitioning were detected since the time the bootloader was configured.
Do you want to proposed bootloader configuration again?**

If yes, all previous bootloader configuration will be lost.
If not, you probably need to change the configuration manually.

Select **OK**. This pop-up window has a count-down timer that selects **OK** automatically if not interrupted with the **Stop** button. The display returns to the **Installation Settings** screen.

- c. Confirm the disk partitions.

A single `root` device and a single `swap` device, both associated with `sda` are the only partitions listed. For example:

```
Partitioning
Create root partition /dev/sda2 (146.92 GB) with ext3
use /dev/sda1 as swap
```

To make additional, site-specific changes, select the **Partitioning** section header and return to the **Expert Partitioner** screen.

23. Confirm the language for the SMW. English (US) is the default language. To change the primary language select the **Language** heading in the **Installation Settings** screen. The **Languages** window opens. Select a language from the drop-down menu. Select multiple secondary languages, if desired. Then select **Accept** at the bottom of the window.
24. On the **Installation Settings** screen, select **Install** at the bottom of the screen. The **Confirm Installation** window appears. To check or change settings, select **Back**; otherwise, select **Install** to confirm and to install the operating system.
The installation runs for approximately 30 minutes. The process automatically reboots the SMW from the hard disk, and the installation continues with system configuration.
25. The **Network Settings** window for configuring the customer network appears. Select the entry labeled **eth0 Customer Network Ethernet**. On the **Overview** tab, select **edit** and do the following:
- Enter the IP address.
 - Enter the subnet mask.
 - Enter the short hostname and select **Next**.
 - Select the **Hostname/DNS** tab. In the **Hostname/DNS and Name Server Configuration** window, enter the hostname, the domain name, the name server values, and the domain names to search. Enter the hostname and domain name separately.

Host Name	Domain Name
smwhost	my.domain.com

If a fully-qualified hostname that includes the domain name is entered, the hostname is accepted but the periods are removed; for example, a hostname of `smwhost.my.domain.com` is converted to `smwhostmydomaincom`.

- Select the **Routing** tab. In the **Routing** window, enter the default gateway IP address. Then select **OK**.

26. The **Clock and Time Zone** window appears.

- Select the appropriate time zone.
- If necessary, adjust the time of day.
- Verify that **Hardware Clock Set To UTC** is selected.

- d. Select **OK**.

At this point, the system finishes booting and enters multiuser mode.

The SMW base operating system, `Cray-SMWbase11SP3`, is now installed.

27. Remove the protective covers from the Fibre Channel (FC) or SAS cables and connectors, clean the ends of the cables and connectors, and reconnect the data cables.
28. Re-seat the previously ejected SMW internal disk drives.
29. Reboot the SMW to allow the SMW to discover the drives properly.

```
smw# reboot
```

30. Use procedure [R815 SMW: Enable an Integrated Dell Remote Access Controller \(iDRAC\)](#) to enable iDRAC on an R815 SMW.

Installation of the base operating system for the R815 SMW is now complete. The system is now ready for installation and configuration of the SMW software. Go to [Install and Configure SMW Software](#) on page 30.

Install and Configure the SMW Software Packages



CAUTION: Shut down the Cray system before installing the Cray SMW software packages.

Install and Configure SMW Software

This procedure takes approximately 30 minutes.

1. Log on to the SMW as `crayadm`, open a terminal window, and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. If the base operating system DVD (`Cray-SMWbase11SP3-`) is still in the CD/DVD drive, eject it.

```
smw# eject
```

3. Place the `Cray 7.2UP04 Software` DVD in the drive and mount it.

```
smw# mkdir -p /media/cdrom
smw# mount /dev/cdrom /media/cdrom
```

NOTE: If problems occur while mounting the DVD after the initial Linux installation, reboot the SMW and perform this procedure again, from the beginning.

4. Copy the `SMWinstall.conf` file from `/media/cdrom` to `/home/crayadm`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
```

5. Change the permissions to make the file writable by the user only.

```
smw# chmod 644 /home/crayadm/SMWinstall.conf
```

6. Edit the `SMWinstall.conf` file to customize it for the installation site. The `SMWinstall.conf` file contains settings for:

- the system interconnection network type, e.g., Aries or Gemini
- the name of the local Network Time Protocol (NTP) servers
- configuring SMWs that have additional disks
- configuring an SMW for a Cray XE6m, Cray XK6m, or Cray XK7m mid-range system
- configuring the Cray Lightweight Log Manager (LLM)
- enabling or disabling the Linux `sar` service on the SMW

```
smw# vi /home/crayadm/SMWinstall.conf
```

- a. Specify the system interconnection network type. There is no default setting.
- b. Adjust LLM settings, as needed.
By default, the `SMWinstall` program enables LLM. To send the logs from the SMW to a site loghost, adjust these settings: `LLM_siteloghost`, `LLM_sitecompatmode`, and `LLM_altrelay`. If the boot RAID controller IP addresses do not start the specified pattern, adjust `llm_raid_ip`.
- c. Leave the Linux `sar` service on the SMW enabled (default), or disable it, as needed.
- d. Leave the controller log forwarding on the SMW enabled (default), or disable it, as needed.
- e. Set the `LOGDISK`, `DBDISK`, and `PMDISK` variables.
The disk device specified by the `LOGDISK` variable must have more disk space than the device specified by the `DBDISK` variable.

1. Use the `LOGDISK` variable to specify the disk device name to be used for logging (`/var/opt/cray/log`, `/var/opt/cray/dump`, and `/var/opt/cray/debug`). The entire disk will be formatted for this use. For the appropriate disk name, see the `LOGDISK` section of the `SMWinstall.conf` file shown below.

IMPORTANT: For an SMW HA system, define this variable as if the system were a stand-alone SMW. The shared storage for logging will be configured later in the SMW HA configuration process.

2. Use the `DBDISK` variable to specify the disk device name to be used for the SMW HSS database (`/var/lib/mysql`). The entire disk will be formatted for this use. For the appropriate disk name for rack-mount models, see the `DBDISK` section of the `SMWinstall.conf` file shown below.

IMPORTANT: For an SMW HA system, define this variable as if the system was a stand-alone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.

3. (Optional) Use the `PMDISK` variable to specify the disk device name to be used for the PostgreSQL Database (`/var/lib/pgsql`). The entire disk will be formatted for this use. For the appropriate disk name for rack-mount models, see the `PMDISK` section of the `SMWinstall.conf` file shown below.

IMPORTANT: For an SMW HA system, do not set the `PMDISK` or `PMMOUNT` variables unless PMDB storage will **not** be shared. These variables are not used on a system that will configure shared PMDB storage as described in [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172 or [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. (The disk device for `/var/lib/pgsql` is set up later in the SMW HA configuration process.) Do not remove the comment character for `PMDISK` and `PMMOUNT` in the `SMWinstall.conf` file.

The *LOGDISK*, *DBDISK*, and *PMDISK* variables must be persistent disk device names. To use the default persistent disk device names, remove the comment character (#) from the relevant lines in the `SMWinstall.conf` file:

```
#For SLES 11 SP3-based rackmount R815 SMWs:
#LOGDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#LOGDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0

#LOGMOUNT=/var/opt/cray/disk/1

#For SLES 11 SP3-based rackmount R815 SMWs:
#DBDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#DBDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0

#DBMOUNT=/var/lib/mysql

#For SLES 11 SP3-based rackmount R815 SMWs:
#PMDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#PMDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0

#PMMOUNT=/var/lib/pgsql
```

For a nonstandard configuration, modify the persistent disk device names accordingly.

- f. For Cray XE6m, Cray XK6m, or Cray XK7m systems: Set the `MCLASS` and `NumChassis` variables. These configuration settings do not apply to Cray XE6m-200, Cray XK6m-200, or Cray XK7m-200 systems.
 1. For a Cray XE6m, Cray XK6m, or Cray XK7m system, set the `MCLASS` variable to `TRUE`. For all other Cray systems, including Cray XE6m-200, Cray XK6m-200, and Cray XK7m-200 systems, `MCLASS` must retain the default setting of `FALSE`.
 2. For a Cray XE6m, Cray XK6m, or Cray XK7m system only, remove the comment character from the `NumChassis` line. The `NumChassis` variable indicates the number of chassis that are in the Cray `MCLASS` system. Choices are 1, 2, 3, 4, 6, 8, 9, 12, 15, or 18.
7. Execute the `SMWinstall` installation script, which updates the base operating system software with SMW security updates and SMW software.

```
smw# /media/cdrom/SMWinstall
```

The output of the installation script is displayed on the console. The `SMWinstall` installation script also creates log files in `/var/adm/cray/logs`.

If for any reason this script fails, it can be rerun without adverse side effects. However, rerunning this script can generate numerous error messages as the script attempts to install already-installed RPMs. Ignore these particular messages.

8. Reboot the SMW.

```
smw# reboot
```

Discover the Cray System Hardware and Power Up the Full System

IMPORTANT: If an installation step fails because of a hardware issue, such as a cabinet failing to power up, resolve that issue and then go back to the last successful step in the installation procedure and continue from there. Do not skip steps or continue out of order.

Bootstrap Hardware Discovery

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. **RESTRICTION:** This step is for a Cray XC Series system Aries™ network SMW only; if installing software on a Cray XE or a Cray XK system (Gemini™ network), skip this step and go to step 3 on page 34.

Update the controller boot image.

The version used in the command argument for `hss_make_default_initrd` should match the version specified in the `lsb-cray-hss` line in output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Wed May 15 12:13:32 CDT 2013 on hssbld0 by bwdev
lsb-cray-hss-7.2.0-1.0702.31509.447
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.31509.447
::: Verifying base RPM list to the manifest
(additional status messages)
::: Removing unwanted files from the root

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.32944.237.
Running hssclone.
Image Clone Complete: /opt/cray/hss-images/default
Running hsspackage.
. . .(additional status messages). . .
inking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img
```

The `xtdiscover` command needs to collect some basic information in order to bootstrap the hardware discovery process. In the next step, be prepared to enter the following information:

```
Abort or continue [a or c]
Network type (g=Gemini, a=Aries, q=quit)
Single-Slot Tester [y or n]
Maximum X cabinet size [1-64]
Maximum Y cabinet size [1-16]
Network topology class [0-3] (0 or 2 for Cray XC30 systems, 0 for Cray XC30-AC
systems)
Boot node name [c0-0c0s0n1]
SDB node name [c0-0c0s2n1] (Cray XE or Cray XK systems)
SDB node name [c0-0c0s1n1] (Cray XC30 systems)
Pathname to the default boot image [/raw0]
```

- The default boot node provided by `xtdiscover` is `c0-0c0s0n1` for Cray XE or Cray XK systems (Gemini™ network) and Cray XC Series systems (Aries™ network).

- The default SDB node provided by `xtdiscover` is `c0-0c0s2n1` for Cray XE or Cray XK systems (Gemini™ network) and Cray XC Series systems (Aries™ network).
- By default, the `xtdiscover` command uses the `/opt/cray/hss/default/etc/xtdiscover.ini` file. During an SMW update, the process will not overwrite the `xtdiscover.ini` file; the new version of the file will be placed in `xtdiscover.ini.dist`.

3. Run the `xtdiscover --bootstrap` command:

```
smw# xtdiscover --bootstrap
***** xtdiscover started *****
Command line: xtdiscover --bootstrap
USER: crayadm pts/0      Jun 14 13:14 (mycomputer.company.com)
Using ini file '/opt/cray/hss/default/etc/xtdiscover.ini'
xtdiscover is about to discover new hardware.
This operation may significantly modify the system database.
Please enter 'c' to continue, or 'a' or 'q' to abort [c]: c
Please enter network type (g=Gemini, a=Aries, q=quit): g
Is this system a Single-Slot Tester? y/n, q=quit [n]: n
Setting system type to Gemini
Discovering Gemini-based system...
Enter maximum X cabinet size (columns) [1-64], q=quit: 1
Enter maximum Y cabinet size (rows) [1-16], q=quit: 1
Adding hosts and routes for 1 cabinet...done.
Enter your system's network topology class [0]: 0 (0 for Cray XC30 systems, 2
for Cray XC30-AC systems)
Setting topology class to 0
Suspending State Manager for discovery phase 1...
Suspend successful.
Saving current configuration...done.
Force new config selected.  Initializing datastore...done.
Discovering cabinets:
[1 out of 1]
Found 1 cabinet.
xtdiscover will create a single system partition (p0)
containing all discovered cabinets.  If you need to create
additional partitions, use 'xtcli part_cfg add'.
Enter the boot node name [c0-0c0s0n1]: c0-0c0s0n1
Enter the SDB node name [c0-0c0s2n1]: c0-0c0s2n1
Enter the absolute pathname to the default boot image [/raw0]: /raw0
Storing base cabinet data...done.
Resuming State Manager for I1 wipe and reboot...Resume successful.
. . .[additional status messages]. . .
Discovery Phase 1 of 3 complete.
```

NOTE: Prior to powering on the cabinets, `xtdiscover` prompts the user to disable any blades that should not be powered on. After disabling these blades, select continue or abort (the `xtdiscover` prompt is `Continue` or `abort [a or c]`). The `xtdiscover` command then proceeds without any further questions.

```
xtdiscover is about to power on the cabinets.
*** IF YOU NEED TO DISABLE COMPONENTS TO AVOID THEM
*** BEING POWERED ON, PLEASE DO SO NOW USING 'xtcli disable'

Please enter 'c' to continue, or 'a' or 'q' to abort [c]: c
Suspending State Manager for discovery phase 2...
Suspend successful.
. . .[additional status messages]. . .
```

Done.

Discovery complete

```
***** xtdiscover finished *****
```

NOTE: If the xtdiscover command fails with the message, The following cabinets were not detected by heartbeat, power cycle the cabinet controller, and retry the xtdiscover -- bootstrap command.

4. **RESTRICTION:** This step is for a Cray XC Series system (Aries™ network) SMW only; if installing software on a Cray XE or a Cray XK system (Gemini™ network), skip this step and go to step 5 on page 35.

Complete the bootstrap process on a Cray XC Series system (Aries™ network) SMW:

- a. Power down the system.

```
smw# xtcli power down s0
```

- b. Reboot the cabinet controllers, then ensure that all cabinet controllers are up.

```
smw# xtccreboot -c all
xtccreboot: reboot sent to specified CCs
smw# xtalive -l cc
```

- c. Power up the system.

```
smw# xtcli power up s0
```

5. **RESTRICTION:** This step is for a Cray XE or a Cray XK system (Gemini network) SMW only. For Cray XC Systems, the bootstrap process is now complete. Continue with Cray system hardware discovery.

Complete the bootstrap process on a Cray XE or a Cray XK system (Gemini network) SMW.

- a. Exit root.

```
smw# exit
```

- b. Log on as `crayadm`, and execute the `xtflash` command to update the L0s and L1s with the current firmware.

```
crayadm@smw> xtflash s0
It took 0 seconds for 'ping' to complete to all L1s.
It took 0 seconds for L1s to become available.
It took 0 seconds for 'ping' to complete to all L0s.
It took 0 seconds for L0s to become available.
spawn fm -qRM -t l1 s0
It took 2 seconds for 'query' to complete to all L1s.
xtrsh -m "-[0-7]$" -f /tmp/xtrsh.Sk5yg9Uz -l root -s "if [ -x /etc/
loadnor ]; \
then /etc/loadnor; fi ; if md5sum /dev/mtd/4 | grep
fec74d6d797dd8b36e68282ad43d0773; \
then echo OLDBIOS; fi"
There is one L1 not up-to-date.
spawn fm -qRM -t l0 s0
It took 2 seconds for 'query' to complete to all L0s.
xtrsh -m s -f /tmp/xtrsh.6ef31uGH -l root -s "if [ -x /etc/loadnor ]; \
then /etc/loadnor; fi ; if md5sum /dev/mtd/4 | grep
fec74d6d797dd8b36e68282ad43d0773; \
```

```

then echo OLDBIOS; fi"
There are 24 L0s not up-to-date.
#####
# Attempt #1 #
#####
spawn fm -RM -t 11 c0-0
It took 157 seconds (2 minutes, 37 seconds) for 'flash' to complete to all
L1s.
There is one L1 not up-to-date.
xtrsh -m "-[0-7]$" -f /tmp/xtrsh.b0zXZERW -l root -s "reboot"
Sleeping for 180 seconds...
It took 0 seconds for 'ping' to complete to all L1s.
It took 0 seconds for L1s to become available.
spawn fm -RM -t 10 c0-0c0s0 c0-0c0s1 c0-0c0s2 c0-0c0s3 c0-0c0s4 c0-0c0s5
c0-0c0s6 c0-0c0s7 \
c0-0c1s0 c0-0c1s1 c0-0c1s2 c0-0c1s3 c0-0c1s4 c0-0c1s5 c0-0c1s6 c0-0c1s7
c0-0c2s0 c0-0c2s1 \
c0-0c2s2 c0-0c2s3 c0-0c2s4 c0-0c2s5 c0-0c2s6 c0-0c2s7
It took 180 seconds (3 minutes, 0 seconds) for 'flash' to complete to all
L0s.
There are 24 L0s not up-to-date.
.
.
.
Everyone is properly flashed.
It took 871 seconds (14 minutes, 31 seconds) to run this application.
crayadm@smw>

```

- If all the L0s are running a compatible BIOS firmware, then xtf1ash completes successfully. Continue to [Discover the Cray System Hardware](#) on page 37.
- If the L0s do not get flashed with the updated BIOS firmware, xtf1ash halts and displays a list of those L0s that need to be updated. These L0s must have their BIOS reflashed before xtf1ash will continue. Perform the BIOS flash only on the components that require it.



CAUTION: Cray recommends reflashing the Cray L0s in groups of a few cabinets at a time rather than reflashing the entire system at once. If there is a catastrophic failure during the BIOS flashing procedure, the only way to recover the BIOS is to reflash each BIOS chip manually. Flashing only a few cabinets at a time limits the risk associated with a catastrophic failure.

IMPORTANT: If the BIOS is flashed on an L1/L0 that already has a compatible BIOS (like an SIO), then the L1/L0 must be power-cycled after being flashed.

In this example, both c0-0c1s3 and c0-0c1s4 need to have their BIOS chips reflashed. Reflash their BIOS chips by using the fm command with the BIOS option -B and a space-separated or comma-separated list of components:

```

crayadm@smw> fm -B -t 10 c0-0c1s3 c0-0c1s4
100% complete
c0-0c1s3: ok
c0-0c1s4: ok

```

With the fm command, specific chassis or cabinets can be specified. For example, to reflash all the L0 BIOS chips on chassis c0-0c1 and cabinets c1-0 and c2-0, enter:

```

crayadm@smw> fm -B -t 10 c0-0c1 c1-0 c2-0

```

- c. Execute the `xtflash` command again after the BIOS chips have been reflashed, to continue reflashing the L1 and L0 Linux firmware.

```
crayadm@smw> xtflash s0
```

The bootstrap process is now complete. The next task is to discover the Cray system hardware.

Discover the Cray System Hardware

To detect the Cray system hardware components on the system, run the `xtdiscover` command. This command creates entries in the system database to describe the hardware. To display the configuration, use the `xtcli` command after running `xtdiscover`. For more detailed information, see the `xtdiscover(8)` man page.

The `xtdiscover` command collects some basic information in order to bootstrap the hardware discovery process, warns that changes will be made, and then confirms whether to abort or continue. Provide the X and Y cabinet sizes as well as topology from [Site-dependent Configuration Values](#) in [Configuration Values](#) on page 22 for `xtdiscover` to accurately discover the hardware. The `xtdiscover` command may prompt the user to execute the `xtbounce` command in a separate window and then continue.

The following steps include an example of the output of the `xtdiscover` command.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. Run the `xtdiscover` command.

```
smw# xtdiscover
***** xtdiscover started *****
Command line: xtdiscover
USER: crayadm pts/1      Sep 10 07:03 (mycomputer.company.com)

Checking system configuration...

Using ini file '/opt/cray/hss/default/etc/xtdiscover.ini'

xtdiscover is about to discover new hardware.
This operation may significantly modify the system database.

Please enter 'c' to continue, or 'a' or 'q' to abort [c]: c

Please enter network type (g=Gemini, a=Aries, q=quit)[a]: a

Setting system type to Aries
Discovering Aries-based system...

Enter maximum X cabinet size (columns) [1-64, last: 1], q=quit: 1
Enter maximum Y cabinet size (rows) [1-32, last: 1], q=quit: 1
Adding hosts and routes for 1 cabinet...done.

A copy of the current partition configuration has been saved in:
    /home/crayadm/hss_db_backup/partitions.09-10-2015.0728

Suspending State Manager for discovery phase 1...
Suspend successful.
Saving current configuration...
Done.
```

```
A backup copy of the HSS database has been saved in:
  /home/crayadm/hss_db_backup/db_backup.09-10-2015.0728.sql

A backup copy of the /etc/hosts file has been saved in:
  /home/crayadm/hss_db_backup/hosts.09-10-2015.0728

Checking current configuration...
Component data received from fstream.

Found 0 existing node power management descriptors
done.

Discovering cabinets:
[1 out of 1]
1 total cabinet.

Gathering base cabinet attributes:
[1 out of 1]
Finished gathering cabinet attributes.

Checking for duplicate MAC addresses...
Done.

Enter your system's network topology class [last: 2]:
Setting topology class to 2

xtdiscover will create a single system partition (p0)
containing all cabinets. If you need to create
additional partitions, use 'xtcli part_cfg add'.

Enter the boot node name [c0-0c0s0n1]:
Enter the SDB node name [c0-0c0s0n2]:
Enter the absolute pathname to the default boot image [/raw0]:

Verifying phase 1 (cabinet) configuration...verification complete.

Storing base cabinet data...done.
Resuming State Manager for power-up and bounce...
Resume successful.

Discovery Phase 1 complete.

xtdiscover is about to power on cabinets.
*** IF YOU NEED TO DISABLE COMPONENTS TO AVOID THEM
*** BEING POWERED ON, PLEASE DO SO NOW USING 'xtcli disable'

Please enter 'c' to continue, or 'a' or 'q' to abort [c]:

Suspending State Manager for discovery phase 2...
Suspend successful.

Loading base component data for discovery phase 2... No updates: skip
done.

Powering on cabinet...
1 cabinet will be powered on:
[1 out of 1]
Cabinets powered on.

Discovering component phase 2 (blade) state:
```

```

[46 out of 46]
Finished discovering component phase 2 (blade) state.

Discovering component phase 2 (blade) attributes:
[46 out of 46]
Finished discovering component phase 2 (blade) attributes.

Verifying phase 2 (blade) configuration...
INFO: 46 blades(s) were modified (component remains, but with changes).

INFO: Modified the following hardware:
    46 slots

INFO: Configuration change details are in
/opt/cray/hss/default/etc/xtdiscover-config-changes.diff

verification complete.

Checking for duplicate MAC addresses...
Done.

Summary of blades discovered:
Total:    48    Service:    7    Empty:    2    Disabled:    0

Discovery Phase 2 complete.

Blades needing to be bounced: 46.

Storing attribute data...done.

Resuming State Manager for bounce...
Resume successful.

46 blades should be bounced using the command
in file /opt/cray/hss/default/etc/xtdiscover-bounce-cmd

Note that this command will also check for certain types
of BIOS errors in the event of a bounce failure.
If this occurs, please follow any additional directions.

In a separate window, please bounce the system now to continue discovery.

```

3. Bounce the system (as `crayadm`) in a separate window.

```
crayadm@smw> /opt/cray/hss/default/etc/xtdiscover-bounce-cmd
```

4. After the `xtbounce` command from the previous step has finished, return to the window in which `xtdiscover` was being run to respond to the remaining prompts, as shown below. Type "c" to continue. And then later type "y" when asked to commit the `xtdiscover` results to the database.

```

After bounce completes, enter 'c' to complete discovery
or 'q' or 'a' to abort [c]: c

Suspending State Manager for discovery phase 3...
Suspend successful.

Discovering component phase 3 (blade/node) attributes:
[46 out of 46]

```

```
Finished discovering component phase 3 attributes.

Discovery Phase 3 complete.

Verifying final configuration...
INFO: 112 total component(s) were added (new; not in DB previously, in any
state).
INFO: 46 total component(s) were modified (component remains, but with changes).

INFO: Added the following hardware:
    14 pdcs (14 enabled, 0 empty or disabled)
    20 accels (20 enabled, 0 empty or disabled)
    56 gpdcs (56 enabled, 0 empty or disabled)
    10 gpdcs (10 enabled, 0 empty or disabled)
    12 hpdcs (12 enabled, 0 empty or disabled)

INFO: Modified the following hardware:
    46 slots

INFO: Configuration change details are in
/opt/cray/hss/default/etc/xtdiscover-config-changes.diff

verification complete.

Discovery complete.

Commit to Database? [y]es, [n]o (repeat stage 3), [a]bort: y

Add discovered power management node descriptors.

Clearing database...done.
Storing component attribute data to database...done.
Storing component attribute data to fast file...done.
Updating component location history...done.

Restarting RSMS daemons for normal operation:
Stopping RSMS services: xtdiagd xtremoted erfbsd xtpmd cm sedc_manager bm nm sm
erdh INFO:
connection to event router lost
INFO: unable to connect to event router, sleeping ...
erd

                                done
Starting RSMS services: erd erdh sm nm bm sedc_manager cm
Waiting for state manager
Starting RSMS services: xtpmd erfbsd xtremoted xtdiagd
Flushing and installing cabinet routes...done.
Partition p0 is active but not booted; skipping background commands.

                                done

Done.

Discovery complete
Waiting for SM ready notification.
Received SM ready notification.
***** xtdiscover finished *****
```

5. **RESTRICTION:** For a Cray XC30 system (Aries network) SMW only. Skip this step (and its substeps) if installing on a Cray XE or XK system.

Update the firmware for a Cray XC30 system:

- a. Exit from `root`.

```
smw# exit
```

- b. Run the `rtr --discover` command to determine the exact configuration of the HSN.

```
crayadm@smw> rtr --discover
```

If the system was not bounced previously, the following message may be displayed:

```
System was not bounced in diagnostic mode, should I re-bounce? Continue (y/n)?
```

If so, enter `y`.

- c. Update the firmware. Execute the `xtzap` command to update the components.

```
crayadm@smw> xtzap -r -v s0
```



CAUTION: The `xtzap` command is normally intended for use by Cray Service personnel only. Improper use of this restricted command can cause serious damage to the computer system.

IMPORTANT: The Cray XC30 system also requires an update to the NVIDIA® BIOS (nvBIOS) for the NVIDIA K20X graphics processing units (GPU). This update is done after CLE has been booted. For more information, see *Installing and Configuring Cray Linux Environment (CLE) Software (S-2444)*.

- d. Run `xtzap` with one or more of the options described below to update if the output of `xtzap` from the previous step includes a "Revision Mismatches" section, indicating some firmware is out of date and needs to be reflashed.

While the `xtzap -a` command can be used to update all components with a single command, it may be faster to use the `xtzap -blade` command when only blade types need to be updated, or the `xtzap -t` command when only a single type needs to be updated. On larger systems, this can be a significant time savings.

This is the list of all cabinet-level components:

```
cc_mc (CC Microcontroller)
cc_bios (CC Tolapai BIOS)
cc_fpga (CC FPGA)
chia_fpga (CHIA FPGA)
```

This is a list of all blade-level components:

```
cbb_mc (CBB BC Microcontroller)
ibb_mc (IBB BC Microcontroller)
anc_mc (ANC BC Microcontroller)
bc_bios (BC Tolapai BIOS)
lod_fpga (LOD FPGA)
node_bios (Node BIOS)
```

```
loc_fpga (LOC FPGA)
qloc_fpga (QLOC FPGA)
```

If the output of the `xtzap` command shows that only a specific type needs to be updated, use the `-t` option with that type (this example uses the `node_bios` type).

```
crayadm@smw> xtzap -t node_bios s0
```

If the output of the `xtzap` command shows that only blade component types need to be updated, use the `-b` option:

```
crayadm@smw> xtzap -b s0
```

If the output of the `xtzap` command shows that both blade- and cabinet-level component types need to be updated, or if there is uncertainty about what needs to be updated, use the `-a` option:

```
crayadm@smw> xtzap -a s0
```

- e. Execute the `xtzap -r -v s0` command again; all firmware revisions should report correctly, except `node_bios`, which will display as "NOT_FOUND" until after the `xtbounce --linktune` command completes.

```
crayadm@smw> xtzap -r -v s0
```

- f. Execute the `xtbounce --linktune` command, which forces `xtbounce` to do full tuning on the system.

```
crayadm@smw> xtbounce --linktune=all s0
```

- g. Execute the `xtzap -r -v s0` command again, to verify that the BIOS version is now correct.

```
crayadm@smw> xtzap -r -v s0
```

6. As user `root`, place the Cray SMW 7.2UP04 Software DVD in the drive and mount it again.

```
smw# mount /dev/cdrom /media/cdrom
```

7. Execute the `SMWconfig` command to configure MySQL.

When prompted for the old root MySQL database password, press the `Enter` key (equivalent to entering an empty string, which is the default password), and create a new root MySQL database password.

NOTE: To become familiar with the requirements for setting and using the root MySQL password, in particular with regard to using the special characters `* ? [< > & ; ! | $`, review the Oracle [MySQL documentation](#).

The `SMWconfig` command can be rerun with no adverse effects. For more detailed information about the `SMWconfig` command, see the `SMWconfig(8)` man page.

```
smw# /media/cdrom/SMWconfig
17:00:33 Date started: Tue Oct 23 17:00:33 2012
17:00:33 Command Line used: /media/cdrom/SMWconfig
17:00:33 Validating command line options
. . .
Please enter your old root MySQL password:
Please confirm your old root MySQL password:
Password confirmed.
Please set your new root MySQL password:
Please confirm your new root MySQL password:
```

```

Password confirmed.
17:00:41 chkconfig rsms on
17:00:41 Checking the status of managers.
17:00:41 Setting MySQL root password.
17:00:41 Beginning to write the external mysql expect script.
17:00:41 Finished writing the external mysql expect script.
17:00:44 Beginning clean up...
17:00:44 Finished cleaning up.
17:00:44 SMWconfig has completed.

```

8. **NOTE:** This step is optional. Perform this step to configure `postfix` on this SMW as a mail transfer agent.

Change the following setting in the `/etc/sysconfig/mail` file on the SMW to prevent the `master.cf` and `main.cf` `postfix` configuration files from being recreated during software updates or fixes.

```
MAIL_CREATE_CONFIG="no"
```

9. Unmount the Cray SMW 7.2UP04 Software DVD.

```
smw# umount /media/cdrom
smw# exit
```

Confirm the SMW is Communicating with System Hardware

The following steps verify that the SMW is functional.

IMPORTANT: For a system configured for SMW high availability (HA) with the SMW failover feature, perform this procedure for the first SMW only. Skip this procedure for the second SMW.

1. Execute the `xthwinv` command on the entire system (`s0`) to examine the hardware inventory and verify that all nodes are visible to the SMW.

```

crayadm@smw> xthwinv s0
Received 24 of 24 responses.
Total number of reported modules: 24
Total number of reported nodes: 86
.
.
.

```

2. Execute the `xtcli status` on the entire system (`s0`).

```

crayadm@smw> xtcli status s0
Network topology: class 0
Network type: Gemini
Nodeid: Service Core Arch| Comp state [Flags]
-----
c0-0c0s0n0: service OP| on [noflags|]
c0-0c0s0n1: service OP| on [noflags|]
c0-0c0s0n2: service OP| on [noflags|]
c0-0c0s0n3: service OP| on [noflags|]
c0-0c0s1n0: - OP| on [noflags|]
c0-0c0s1n1: - OP| on [noflags|]
c0-0c0s1n2: - OP| on [noflags|]
c0-0c0s1n3: - OP| on [noflags|]

```

```

c0-0c0s2n0: service OP| on [noflags|]
c0-0c0s2n1: service OP| on [noflags|]
c0-0c0s2n2: service OP| on [noflags|]
c0-0c0s2n3: service OP| on [noflags|]
c0-0c0s3n0: - OP| on [noflags|]
c0-0c0s3n1: - OP| on [noflags|]
c0-0c0s3n2: - OP| on [noflags|]
c0-0c0s3n3: - OP| on [noflags|]
.
.
.
-----

```

- Execute the `rtr -R` commands on the entire system (`s0`) to ensure that the SMW is functional. Note that the `rtr -R` command produces no output unless there is a routing problem.

```
crayadm@smw> rtr -R s0
```

- Execute the `xtmcinfo -t -u` command to retrieve microcontroller information from cabinet control processors and blade control processors.

NOTE: In this example, the `Contents of timestamp...` line appears only in output from Cray XE systems, while output from Cray XC systems contains only the `How long have they been up...` section.

```

crayadm@smw> xtmcinfo -t -u s0
Contents of timestamp...
c0-0c0s0 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s1 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s2 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s3 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s4 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s5 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s6 - Wed Apr 20 01:27:27 CDT 2012
c0-0c0s7 - Wed Apr 20 01:27:27 CDT 2012
.
.
.
c0-0c2s6 - Wed Apr 20 01:27:27 CDT 2012
c0-0c2s7 - Wed Apr 20 01:27:27 CDT 2012
How long have they been up...
c0-0c0s0 - 14:11:49 up 23 min, load average: 0.00, 0.01, 0.03
c0-0c0s1 - 14:11:49 up 23 min, load average: 0.53, 0.13, 0.06
c0-0c0s2 - 14:11:50 up 23 min, load average: 0.00, 0.00, 0.02
c0-0c0s3 - 14:11:50 up 23 min, load average: 0.14, 0.03, 0.01
c0-0c0s4 - 14:11:50 up 23 min, load average: 0.00, 0.01, 0.03
c0-0c0s5 - 14:11:50 up 23 min, load average: 0.01, 0.10, 0.12
c0-0c0s6 - 14:11:50 up 23 min, load average: 0.04, 0.10, 0.10
c0-0c0s7 - 14:11:50 up 23 min, load average: 0.01, 0.07, 0.08

```

Change the Default SMW Passwords

After completing the installation, change the default SMW passwords on the SMW. The SMW contains its own `/etc/passwd` file that is separate from the password file for the rest of the system. To change the passwords on the SMW, log on to the SMW as `root` and execute the following commands:

```
crayadm@smw> su - root
smw# passwd root
smw# passwd crayadm
smw# passwd cray-vnc
smw# passwd mysql
```

For rack-mount SMWs (both R815 and R630 models), it is also necessary to change the default iDRAC password. See [Change the Default iDRAC Password](#).

Set Up the SUSE Firewall and IP Tables

The SMW software includes a firewall. The following steps enable and configure the firewall.

TIP: It is not necessary to shut down the system before performing this task.

1. Before modifying the SUSE Firewall settings, make a copy of the configuration file:

```
smw# cp -p /etc/sysconfig/SuSEfirewall12 /etc/sysconfig/SuSEfirewall12.orig
```

2. Using the SuSEfirewall12 program and the following steps, change the IP tables rules to close off all unnecessary ports on the SMW.

```
smw# iptables -L
smw# vi /etc/sysconfig/SuSEfirewall12
```

Change the settings of these variables to the values shown:

```
FW_DEV_EXT="any eth0"

FW_DEV_INT="eth1 eth2 eth3 eth4 lo"

FW_SERVICES_EXT_UDP="161"

FW_TRUSTED_NETS="your_bootnode_ipaddress,tcp,7004 \
    your_syslognode_ip,udp,514 your_sdbnode_ip,tcp,6811:6815"
```

For example:

```
smw# diff /etc/sysconfig/SuSEfirewall12.orig /etc/sysconfig/SuSEfirewall12
99c99
< FW_DEV_EXT="eth-id-00:30:48:5c:b0:ee eth0"
---
> FW_DEV_EXT="any eth0"
114c114
< FW_DEV_INT="eth-id-00:0e:0c:b4:df:64 eth-id-00:0e:0c:b4:df:65
    eth-id-00:0e:0c:b4:df:66 eth-id-00:0e:0c:b4:df:67 eth1 eth2 eth3 eth4"
---
> FW_DEV_INT="eth1 eth2 eth3 eth4 lo"
263c263
< FW_SERVICES_EXT_UDP=""
---
> FW_SERVICES_EXT_UDP="161"
394c394
< FW_TRUSTED_NETS=""
---
> FW_TRUSTED_NETS="10.3.1.254,tcp,7004 10.5.1.2,udp,514 10.5.1.2,tcp,6811:6815"
```

NOTE: 10.3.1.254 is the boot node's IP address for eth0 on the network between the boot node and the SMW. 10.5.1.2 is used for the DRBD connection on SMW HA, but is different for the first SMW and the second SMW: the first is 10.5.1.2 (active) and the second is 10.5.1.3 (passive).

3. Invoke the modified configuration.

```
smw# /etc/init.d/SuSEfirewall2_init start
smw# /etc/init.d/SuSEfirewall2_setup start
```

4. Execute the following commands to start the firewall at boot time.

```
smw# chkconfig SuSEfirewall2_init on
smw# chkconfig SuSEfirewall2_setup on
```

5. Verify the changes to the iptables.

```
smw# iptables -nvL
```

SSH access is one of the protocols permitted through the firewall from the external network to the SMW. For information about how to use Virtual Network Computing (VNC) through an SSH tunnel, see [Enable Remote Access to the SMW using VNC](#).

Configure the Simple Event Correlator (SEC)

The System Management Workstation (SMW) 7.2.UP04 release includes the Open Source simple event correlator (SEC) package, `sec-2.7.0`, and an SEC support package, `cray-sec-version`. The SEC support package contains control scripts to manage the starting and stopping of SEC around a Cray mainframe boot session, in addition to other utilities.

To use the Cray SEC, see *Configure SEC Software (S-2542)* for configuration procedures.

Finish Installing SMW Software on the First SMW

Prerequisites

Before using this procedure, ensure that the operating system and Cray SMW software is correctly installed on the first SMW.

Use these HA-specific steps to finish the SMW software installation on the first SMW.

1. If you changed the default root password on the SMW, you must also change the default iDRAC password to the same password. For more information, see [Passwords For an SMW HA System](#) on page 18.

IMPORTANT: Both SMWs must have exactly the same password for the root and iDRAC accounts.

2. Configure email on the SMW. The SMW HA system uses email for failover notification. For information about configuring email on an SMW, see http://www.postfix.org/BASIC_CONFIGURATION_README.html.
3. Ensure that the boot RAID is connected before you continue to the next procedure.

Configure the Boot RAID for SMW HA

After installing the SMW software on the first SMW, create LUNS for the shared storage on the boot RAID. The SMW HA system requires LUNS for the MySQL database, log directory, and /home. To determine the space needed for shared storage, see [Shared Storage on the Boot RAID](#) on page 8 and [Minimum Boot RAID LUN Sizes for SMW Failover](#) on page 15.

If shared storage on the boot RAID will be used for the power management database (PMDB), create an additional LUN with sufficient space for the PMDB. Ensure that RAID disk has enough space to hold all data for the power management database (PMDB) by checking the size of /var/lib/pgsql. Use the following command to display the size of the local data:

```
smw1:~ # du -hs /var/lib/pgsql
```

IMPORTANT: Mirrored storage is preferred for the PMDB (see [Storage for the Power Management Database \(PMDB\)](#) on page 9).

Use the following procedures to configure and zone the boot RAID, including the required LUNs for the SMW HA cluster. Use the recommended boot RAID LUN configuration for a single SMW.

IMPORTANT:

- Any existing data on the boot RAID disks will be wiped out during installation.
- Proceed with care! Make sure to use the correct disk names.

Configure the Boot RAID

NOTE: Cray ships systems with much of this software installed and configured. Performing all of the steps in the following boot RAID procedures may not be necessary unless the configuration needs to be changed.

In typical system installations, the RAID provides the storage for both the boot node root file systems and the shared root file system. Although the boot node manages these file systems during normal operation, the SMW performs the initial installation of the CLE operating system and the Cray software packages on the boot RAID disks.

For a system configured for SMW high availability (HA) with the SMW failover feature, the boot RAID is also used for the shared log, MySQL database, and /home file system.

In typical system installations, RAID units provide user and scratch space and can be configured to support a variety of file systems. For more information about configuring RAID, see [Manage Lustre for the Cray Linux Environment \(CLE\) \(S-0010\)](#), which is provided with the CLE release package.

Cray provides support for system boot RAID from two different vendors: Data Direct Networks, Inc. (DDN) and NetApp, Inc. To configure the boot RAID, use the procedures for the type of boot RAID at this site:

DDN boot RAID [Configure and Zone the Boot RAID for a DDN Storage System](#) on page 49

Includes procedures to configure and zone LUNs (logical unit numbers) using either te1net or the Silicon Storage Appliance Manager GUI.

**NetApp, Inc.
Engenio™ boot
RAID**

[Configure the Boot RAID for a NetApp, Inc. Engenio Storage System](#) on page 58

Includes procedures to install the SANtricity Storage Manager utility, which is then used to create and configure volumes and then assign them to LUNs.

After configuring the boot RAID, use the following procedures, as applicable:

- If a QLogic Fibre Channel Switch is used, follow the procedures in [Zone the QLogic™ FC Switch](#) on page 62.
- To rediscover the LUNs (logical unit numbers) and zones that have been created, use this procedure: [Reboot the SMW and Verify LUNs are Recognized](#) on page 64.
- To partition the LUNs, use this procedure: [Partition the LUNs](#) on page 64.
- For systems that use the CLE feature Direct-attached Lustre (DAL), use this procedure: [Create LUNs and an IMPS File System for DAL](#).

Prerequisites for these procedures:

- The SMW has an Ethernet connection to the Hardware Supervisory System (HSS) network.
- The boot nodes have Ethernet connections to the SMW.
- The SMW has a Fibre Channel (FC) or Serial Attached SCSI (SAS) connection to the boot RAID or to an FC switch.
- The boot nodes have an FC or SAS connection to the boot RAID or to an FC switch.
- The SDB node has an FC or SAS connection to the boot RAID or to an FC switch. The SDB node may have an Ethernet connection.
- If a dedicated `syslog` node is configured, it has an FC or SAS connection to the boot RAID or to an FC switch. A dedicated `syslog` node does not have an Ethernet connection.

Prerequisites and Assumptions for Configuring the Boot RAID

NOTE: Cray ships systems with much of this software installed and configured. You may not need to perform all of the steps described in this chapter unless you are making changes to the configuration.

Cray provides support for system boot RAID from two different vendors: Data Direct Networks, Inc. (DDN) and NetApp, Inc. If you are configuring a DDN boot RAID, follow the procedures in [Configure and Zone the Boot RAID for a DDN Storage System](#) on page 49. If you are configuring a NetApp, Inc. Engenio™ boot RAID, follow the procedures in [Configure the Boot RAID for a NetApp, Inc. Engenio Storage System](#) on page 58.

In typical system installations, the RAID provides the storage for both the boot node root file systems and the shared root file system. Although the boot node manages these file systems during normal operation, the SMW performs the initial installation of the CLE operating system and the Cray software packages on the boot RAID disks.

For a system configured for SMW high availability (HA) with the SMW failover feature, the boot RAID is also used for the shared log, MySQL database, and `/home` file system.

In typical system installations, RAID units provide user and scratch space and can be configured to support a variety of file systems. For more information about configuring RAIDs, see *Managing Lustre for the Cray Linux Environment (CLE)* (S-0010), which is provided with your CLE release package.

Prerequisites for this task are:

- The SMW has an Ethernet connection to the Hardware Supervisory System (HSS) network.
- The boot nodes have Ethernet connections to the SMW.

- The SMW has a Fibre Channel (FC) or Serial Attached SCSI (SAS) connection to the boot RAID or to a FC switch.
- The boot nodes have a FC or SAS connection to the boot RAID or to a FC switch.
- The SDB node has a FC or SAS connection to the boot RAID or to a FC switch.
- If a dedicated `syslog` node is configured, it has a FC or SAS connection to the boot RAID or to a FC switch.
- The DDN RAID uses LUNs (Logical Units). The NetApp, Inc. RAID uses volumes.
The SDB node may have an Ethernet connection. A dedicated `syslog` node does not have an Ethernet connection.

Configure and Zone the Boot RAID for a DDN Storage System

NOTE: The instructions in the procedures listed below supersede the documentation supplied by the RAID manufacturer.

- To configure a DDN storage system using the command-line interface (`telnet`), use the following procedures:
 1. [Change the Boot RAID Password for DDN Devices using telnet](#) on page 49
 2. [Configure the LUNs for DDN Devices using telnet](#) on page 50
 3. [Zone the LUNs for DDN Devices using telnet](#) on page 51
- To configure a DDN storage system using the DDN Silicon Storage Appliance (S2A) Manager GUI:
 1. Install the GUI software using these procedures:
 - a. [Install the Silicon Storage Appliance Manager Software for DDN Devices](#) on page 54
 - b. [Identify the Installed Version of the DDN Silicon Storage Appliance Manager Software](#) on page 55
 2. Change the password. The password should be changed on all DDN boot RAIDs, whether using `telnet`, as described in procedure 2 above, or using the S2A Manager GUI. See DDN documentation on the S2A Manager for instructions on how to change the password using the GUI.
 3. Configure and zone LUNs using this procedure: [Configure and Zone the LUNs for DDN Devices using the S2A Manager GUI](#) on page 57

To enable and configure remote logging of DDN messages so that the controller's internal error log does not overflow, use this procedure : [Configure Logging using Syslog on DDN Storage Devices](#) on page 57. This is a `telnet` procedure. DDN documentation for the S2A Manager may have instructions on how to do this using the GUI.

Configure a DDN Storage System for the Boot RAID via `telnet`

Change the Boot RAID Password for DDN Devices using `telnet`

As part of configuring the boot RAID using DDN, Cray recommends changing the passwords for `admin` and `user`.

The RAID default administrative login name is `admin` and the default password is `password`. The standard IP address for the RAID controller is `10.1.0.1`. The secondary RAID controller, if used, is `10.1.0.2`.

1. If necessary, log on to the SMW as `crayadm`.
2. From the SMW, use the `telnet` command to log on to the RAID controller.

```
crayadm@smw> telnet 10.1.0.1
```

3. Log on as `admin` and enter the default password.

```
login: admin
Password:
CAB01-L1:
```

4. Change the password by entering the current password for `admin`, and then enter the information as prompted.

```
CAB01-L1: password
Enter current password:
*****
Enter a new name to replace 'admin', or return to leave unchanged:

Administrative user name 'admin' unchanged.

Enter new password:
*****
Re-enter the new password:
*****

Enter a new name to replace 'user', or return to leave unchanged:

General user name 'user' unchanged.

Enter new password:
*****
Re-enter the new password:
*****

Password for general user 'user' unchanged.

Committing changes.
CAB01-L1:
```

Configure the LUNs for DDN Devices using telnet

Configure the boot RAID with a minimum of six LUNs to support the various system management file systems. Cray recommends nine LUNs, configured as described in "Configuring the Boot RAID LUNs or Volume Groups" in *CLE Installation and Configuration Guide (S-2444)*.

Three LUNs are required for High Availability SMW. If the CLE feature Direct Attached Lustre (DAL) is configured on SMW HA, another LUN is needed for Image Management and Provisioning System (IMPS) data.

1. On the RAID controller, run the `lun` command to view the existing configuration. If the RAID is already configured, the command returns the current LUN status.

```
CAB01-L1: lun
```

Logical Unit Status

LUN	Label	Owner	Status	Capacity (Mbytes)	Block		
					Size	Tiers	Tier list

- Run the `lun delete` command to delete preexisting LUNs.

```
CAB01-L1: lun delete=x
```

where *x* is the LUN to be deleted.

- Run the `lun add` command to add, configure, and format a LUN.

```
CAB01-L1: lun add=x
```

where *x* is the LUN to be configured. This command initiates a dialog similar to the following example.

```
Enter the LUN (0..127) to add, or 'e' to escape: 0
  Enter a label for LUN 0 (up to 12 characters): bootroot0
  You can create a single LUN or a LUN group
  of smaller LUNs with equal capacity.
  Do you want to create a LUN group? (y/N): n
  Enter the capacity (in Mbytes) for LUN 0, LUN0
  0 for all available capacity (default): 40000
  Enter the number of tiers (1..8)
  Default will auto select, 'e' to escape: 1
  Enter the tiers, each one on a new line, or 'e' to escape: 1
  Enter the block size, (512, 1024, 2048, 4096, 8192)
  Default is 512, 'e' to escape: 4096
  Operation successful: LUN 0 was added to the system.
  The LUN must be formatted before it can be used.
  Would you like to format the LUN now? (y/N): y
```

Repeat this step for each LUN to be configured and formatted.

The recommended boot RAID LUN configuration is shown in the following table from *CLE Installation and Configuration Guide (S-2444)*, which is provided with the CLE release package.

Table 8. Recommended Boot RAID LUN Values

LUN	Label	Size (1-50 Cabinets)	Size (50+ Cabinets)	Segment Size
0	bootroot0	40GB	70GB	256KB
1	shroot0	280GB	370GB	256KB
2	sdb0	60GB	80GB	256KB
3	bootroot1	40GB	70GB	256KB
4	shroot1	280GB	370GB	256KB
5	sdb1	60GB	80GB	256KB
6	bootroot2	40GB	70GB	256KB
7	shroot2	280GB	370GB	256KB
8	sdb2	60GB	80GB	256KB

NOTE: On systems that are configured for SMW HA with the SMW failover feature, the boot RAID also includes space for the shared log, MySQL database, and /home file system.

Zone the LUNs for DDN Devices using `teInet`

After the LUNs are configured and formatted, grant host access to the LUNs using a process called *zoning*. Zoning maps a host port on the RAID controller to the LUNs that the host accesses. Although it is possible to allow all hosts to have access to all LUNs, Cray recommends that each host be granted access only to the LUNs it requires.

NOTE: If a LUN is to be shared between failover host pairs, give each host access to the LUN. The SMW host port should be given access to all LUNs.

All zoning relationships are defined with the `zoning` command.

1. Display the current zoning summary.

```
CAB01-L1: zoning
```

```
Port Zoning Summary:
          LUN Zoning
Port  World Wide Name  (External LUN, Internal LUN)
-----
  1    21000001FF030759
  2    22000001FF030759
  3    23000001FF030759
  4    24000001FF030759
```

If the LUNs are not zoned, this command displays only the Ports and their World Wide Names, the unique identifier in the Fibre Channel storage network.

If the LUNs are zoned, consider clearing some or all of the current settings.

2. Execute the `zoning edit=x` command to begin zoning the LUNs, where *x* is the Port number.

```
CAB01-L1: zoning edit=1
```

This command returns the following prompt:

```
Enter the new LUN zoning for host port 1.
```

```
Enter the unique LUN mapping, as follows:
```

```
G.1  GROUP.LUN number
P    Place-holder
R    Before GROUP.LUN to indicate Read-Only
N    Clear current assignment
<cr> No change
E    Exit command
?    Display detailed help text
```

```
External LUN 0: is not mapped.  Enter new internal LUN:
```

3. The zoning dialog cycles through each LUN in sequence. Enter the LUN number to map the LUN to the selected port, press the Enter key to leave the LUN mapping unchanged, or enter `n` to clear (remove) the current LUN mapping. When finished mapping LUNs to the selected port, enter `e` to exit. The updated zoning summary displays.

For Port 1, map LUNs 0 through 2 as follows:

```
External LUN 0: is not mapped.  Enter new internal LUN: 0
External LUN 1: is not mapped.  Enter new internal LUN: 1
```

```
External LUN 2: is not mapped. Enter new internal LUN: 2
External LUN 3: is not mapped. Enter new internal LUN: e
*** Host Port 1: zoning has been updated! ***
```

```

                                LUN Zoning
      Port  World Wide Name      (External LUN, Internal LUN)
-----
      1     21000001FF020320      000,000    001,001    002,002
      2     22000001FF020320
      3     23000001FF020320
      4     24000001FF020320

```

For Port 2, map LUNs 3 through 5, as follows:

```
External LUN 0: is not mapped. Enter new internal LUN: n
External LUN 1: is not mapped. Enter new internal LUN: n
External LUN 2: is not mapped. Enter new internal LUN: n
External LUN 3: is not mapped. Enter new internal LUN: 3
External LUN 4: is not mapped. Enter new internal LUN: 4
External LUN 5: is not mapped. Enter new internal LUN: 5
External LUN 6: is not mapped. Enter new internal LUN: e
*** Host Port 2: zoning has been updated! ***
```

```

                                LUN Zoning
      Port  World Wide Name      (External LUN, Internal LUN)
-----
      1     21000001FF020320      000,000 001,001 002,002
      2     22000001FF020320      003,003 004,004 005,005
      3     23000001FF020320
      4     24000001FF020320

```

For Port 3, map LUNs 6 through 8, as follows:

```
External LUN 0: is not mapped. Enter new internal LUN: n
External LUN 1: is not mapped. Enter new internal LUN: n
External LUN 2: is not mapped. Enter new internal LUN: n
External LUN 3: is not mapped. Enter new internal LUN: n
External LUN 4: is not mapped. Enter new internal LUN: n
External LUN 5: is not mapped. Enter new internal LUN: n
External LUN 6: is not mapped. Enter new internal LUN: 6
External LUN 7: is not mapped. Enter new internal LUN: 7
External LUN 8: is not mapped. Enter new internal LUN: 8
External LUN 9: is not mapped. Enter new internal LUN: e
*** Host Port 3: zoning has been updated! ***
```

```

                                LUN Zoning
      Port  World Wide Name      (External LUN, Internal LUN)
-----
      1     21000001FF020320      000,000 001,001 002,002
      2     22000001FF020320      003,003 004,004 005,005
      3     23000001FF020320      006,006 007,007 008,008
      4     24000001FF020320

```

4. Enable continuous LUN verification. This process runs in the background with a performance penalty of approximately 1%.

```
CAB01-L1: lun verify=on
[...]
Please enter a LUN ('a' for all LUNs, 'q' to quit): a
All valid LUNs selected
Do you want the verify to run continuously? (y/N): y
```



CAUTION: Turn off LUN verification before performing maintenance on disk subsystems. Running LUN verification and swap or moving a back-end channel cable could disrupt an entire channel of drives.

- The final LUN zoning should look like the following example. If it does not, edit any ports using the procedure previously described.

```
CAB01-L1: zoning
```

Port	World Wide Name	LUN Zoning		
		(External LUN, Internal LUN)		
1	21000001FF020320	000,000	001,001	002,002
2	22000001FF020320	003,003	004,004	005,005
3	23000001FF020320	006,006	007,007	008,008
4	24000001FF020320			

- When finished zoning the LUNs, close the `telnet` connection to the RAID controller and return to the SMW.

```
CAB01-L1: logout
```

- If not already logged on as `root`, `su` to `root`.

```
crayadm@smw> su - root
```

- Reboot the SMW. This enables the SMW to recognize the new LUN configuration and zoning information.

```
smw# reboot
```

After finishing creating, formatting, and zoning the LUNs on the boot RAID, partition them. Although partitioning occurs on the SMW, use the boot LUN partition table included in *CLE Installation and Configuration Guide* (S-2444), which is provided with the CLE release package, because recommendations may change with each CLE release.

Install the Silicon Storage Appliance Manager GUI

Optionally, you can use the Silicon Storage Appliance Manager GUI to configure a DDN storage system for your Boot RAID.

Install the Silicon Storage Appliance Manager Software for DDN Devices

This procedure describes how to install the DDN Silicon Storage Appliance (S2A) Manager on the SMW. The S2A Manager has a graphical user interface, which can be used as an alternative to `telnet` to configure and zone the LUNs for DDN storage devices.

- Log on to the SMW and enable X windows port forwarding.
 - If already logged on to the SMW, `su` to `root` and enter the following command:

```
smw# ssh -X
```

- If not logged on to the SMW, enter the following command:

```
workstation> ssh -X smw
```

- Copy (as `root` user) the Silicon Storage Appliance Manager installation file (`directGUI_version_Linux_x64_NoVM.bin`) to a temporary location on the SMW, such as `/tmp`. The Silicon Storage Appliance Manager installation file is on a CD that is provided with the system.

- Move to the directory that the installation file was copied to.

```
smw# cd /tmp
```

- Initiate the installation process:

```
smw# ./directGUI_version_Linux_NoVM.bin
```

The **Introduction** window displays.

- Select **Next**.

The **License Agreement** window displays.

- Accept the License Agreement and select **Next**.

The **Choose Uninstall Folder** window displays, showing the installed software.

- Select the uninstall folder (file path) to uninstall and select **Next**.

- Select the **complete** icon for a complete uninstall.
- Select **Done**.

The **Choose Java Virtual Machine** window opens.

- Choose a Java Virtual Machine and select **Next**.

The **Choose Install Folder** window opens.

- Choose the install folder and select **Next**.

The **Choose Link Folder** window opens.

- Choose the link folder and select **Next**.

The **Pre-Installation Summary** window opens.

- Review the pre-installation summary. If everything is correct, select **Install**.

When the installation completes, the **Install Complete** window opens, indicating the installation is complete.

- Select **Done**.

Identify the Installed Version of the DDN Silicon Storage Appliance Manager Software

Prerequisites

The user must be logged in as `root` to perform this task.

- Start the GUI using one of the following:

```
smw# /usr/local/ddn/Silicon_Storage_Appliance_Manager
```

```
smw# /opt/Silicon_Storage_Appliance_Manager/Silicon_Storage_Appliance_Manager
```

2. Select from **Help > About** on the Silicon Storage Appliance Manager menu bar to identify the version of the GUI. If using a Silicon Storage Appliance Manager GUI version earlier than 2.07, download and install the new Silicon Storage Appliance Manager software.

Install the Silicon Storage Appliance Manager Software for DDN Devices

This procedure describes how to install the DDN Silicon Storage Appliance (S2A) Manager on the SMW. The S2A Manager has a graphical user interface, which can be used as an alternative to `telnet` to configure and zone the LUNs for DDN storage devices.

1. Log on to the SMW and enable X windows port forwarding.
 - If already logged on to the SMW, `su` to `root` and enter the following command:

```
smw# ssh -X
```

- If not logged on to the SMW, enter the following command:

```
workstation> ssh -X smw
```

2. Copy (as `root` user) the Silicon Storage Appliance Manager installation file (`directGUI_version_Linux_x64_NoVM.bin`) to a temporary location on the SMW, such as `/tmp`. The Silicon Storage Appliance Manager installation file is on a CD that is provided with the system.
3. Move to the directory that the installation file was copied to.

```
smw# cd /tmp
```

4. Initiate the installation process:

```
smw# ./directGUI_version_Linux_NoVM.bin
```

The **Introduction** window displays.

5. Select **Next**.

The **License Agreement** window displays.

6. Accept the License Agreement and select **Next**.

The **Choose Uninstall Folder** window displays, showing the installed software.

7. Select the uninstall folder (file path) to uninstall and select **Next**.

- a. Select the **complete** icon for a complete uninstall.

- b. Select **Done**.

The **Choose Java Virtual Machine** window opens.

8. Choose a Java Virtual Machine and select **Next**.

The **Choose Install Folder** window opens.

9. Choose the install folder and select **Next**.

The **Choose Link Folder** window opens.

10. Choose the link folder and select **Next**.

The **Pre-Installation Summary** window opens.

11. Review the pre-installation summary. If everything is correct, select **Install**.

When the installation completes, the **Install Complete** window opens, indicating the installation is complete.

12. Select **Done**.

Configure and Zone the LUNs for DDN Devices using the S2A Manager GUI

Use the DDN Silicon Storage Appliance (S2A) Manager to configure and zone LUNs. See the DDN S2A Manager GUI documentation for instructions.

Configure Remote Logging of DDN Messages

To avoid overflowing the controller's internal error log, follow this procedure to enable and configure remote logging of DDN messages.

Configure Logging using Syslog on DDN Storage Devices

The message logs for DDN S2A 8500 and S2A 9550 storage devices are written to a fixed area of controller memory. Some storage problems can produce output that exceeds the amount of available memory space in the controller and force initial error messages, which are necessary to isolate the root cause of the problem, to overflow and scroll out of memory.

Cray recommends reconfiguring the S2A network logging to enable the SMW to store and retrieve failure information.

When a RAID controller is configured to send syslog messages to the SMW, the messages from the RAID are automatically stored in `/var/log/cray/log/raid-yyyymmdd`, assuming the RAID controllers use an IP address beginning with `10.1.0`. If the RAID controllers begin with a different IP address, set the `llm_raid_ip` variable in the `SMWinstall.conf` file, remount the SMW installation media, and rerun the SMW installer with the `forceupdate` option.

```
smw# vi /home/crayadm/SMWinstall.conf
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37275.648-1.iso /media/cdrom
smw# /media/cdrom/SMWinstall --forceupdate
```

1. Enable syslog on the DDN device to the SMW.

Log on as `admin` to the DDN device. This example uses the command-line interface (`telnet`) and uses the boot RAID and its default IP address.

NOTE: Use the password that was set in [Change the Boot RAID Password for DDN Devices using telnet](#) on page 49.

```
smw# telnet 10.1.0.1
login: admin
Password: ****
```

Verify the current network settings by using the `network` command. The settings for `syslog` appear at the end of the display.

```

CAB01-L1: network
...

Syslog:                DISABLED
Syslog IP Address:    0.0.0.0
Syslog Port unit #1: 514
Syslog Port unit #2: 514

```

Enable syslog to the IP address 10.1.1.1 for the SMW eth1 interface on the default port of 514.

```

CAB01-L1: network syslog=on
CAB01-L1: network syslogip=10.1.1.1
CAB01-L1: network syslogport=514

```

Verify the new settings.

```

CAB01-L1: network
...

Syslog:                ENABLED
Syslog IP Address:    10.1.1.1
Syslog Port unit #1: 514

```

2. Log off the DDN device.

```

CAB01-L1: logout

```

Configure the Boot RAID for a NetApp, Inc. Engenio Storage System

To configure the boot RAID for a NetApp, Inc. Engenio Storage System, it is necessary to first install the SANtricity Storage Manager Utility. Then, the rest of the procedures use that utility to create and configure volumes, assign them to LUNs, and configure remote logging.

NOTE: The instructions in these procedures apply for both SAS (Serial Attached SCSI) and Fibre Channel RAIDs and supersede the documentation supplied by the RAID manufacturer.

1. Install the SANtricity Storage Manager Utility using this procedure: [Install SANtricity Storage Manager Software for NetApp, Inc. Engenio Devices](#) on page 58. SANtricity is provided as a separate package and is installed from a CD. It may already be installed on the SMW.
2. Use the SANtricity Storage Manager utility from NetApp, Inc. to perform the remaining procedures. These procedures assume familiarity with using the SANtricity interface.
 - a. To create the volume group, use this procedure: [Create the Boot RAID Volume Group for NetApp, Inc. Engenio devices](#) on page 60.
 - b. To create the LUNs within the volume group, use this procedure: [Create and Configure Volumes for NetApp, Inc. Engenio Devices](#) on page 61.
 - c. To configure remote logging of the boot RAID messages, use this procedure: [Configure Remote Logging of NetApp, Inc. Engenio Storage System Boot RAID Messages](#) on page 62.

Install SANtricity Storage Manager Software for NetApp, Inc. Engenio Devices

The SANtricity Storage Manager software is generally preinstalled and the SANtricity media is shipped with the system. If the SANtricity software is installed, then the SMclient executable will be found in /opt/SMgr/client/SMclient. If this Cray system does not have the software installed on the SMW, then install it using the procedure in [Install the SANtricity Software](#) on page 59.

Install the SANtricity Software

1. Log on to the SMW as `root`.

```
crayadm@smw> su - root
```

2. Install SANtricity Storage Manager from the CD or from a directory.
 - To install from the SANtricity Storage Manager CD, insert it into the SMW CD drive. Verify that the media has mounted automatically; if not, mount it manually.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install from the SMIA-LINUX-10.70.A0.25.bin file, copy SMIA-LINUX-10.70.A0.25.bin to /home/crayadm.

```
smw# cp ./SMIA-LINUX-10.70.A0.25.bin /home/crayadm/
```

3. Set the `DISPLAY` environment variable.

```
smw# export DISPLAY=:0.0
```

4. Verify that the X Window System is functioning by launching `xterm` or executing the `xlogo` utility.

```
smw# xterm
```

or:

```
smw# xlogo
```

Then, exit the Xlogo window or `xterm`.

5. Invoke the executable file.

If installing from the CD:

```
smw# /bin/bash /media/cdrom/install/SMIA-LINUX-10.70.A0.25.bin
```

If installing from a directory:

```
smw# /home/crayadm/SMIA-LINUX-10.70.A0.25.bin
```

6. Select **Next**. The **License Agreement** window displays.
7. Accept the license agreement and select **Next**. The **Select Installation Type** window displays.
8. Select **Typical (Full Installation)**, then select **Next**.
The **Multi-Pathing Driver Warning** window displays.
9. Select **OK**. The **Pre-Installation Summary** window displays.
10. Select **Install**.
The **Installing SANtricity** window displays and shows the installation progress. When the installation completes, an **Install Complete** window appears.
11. Select **Done**. The SANtricity client is installed in `/usr/bin/SMclient` and is currently running.

12. To execute SMclient from the `crayadm` account, change the ownership and permissions for executable files; otherwise, execute SMclient as `root`.

```
smw# cd /opt
smw# chown crayadm SMgr
smw# chmod 775 SMgr
smw# cd SMgr/client
smw# chmod 755 SMcli SMclient
smw# cd /var/opt
smw# chown -R crayadm:crayadm SM
smw# chmod -R ug+w SM
```

13. Close the file browser and eject the CD.

```
smw# eject
```

Configure the LUNs for NetApp, Inc. Engenio Devices

Create the Volume Group and the LUNs that are members of it.

Create the Boot RAID Volume Group for NetApp, Inc. Engenio devices

Create the 3+1 Volume Group and 1 Global Hot Spare across the first five disks for a 4.5 TB Volume Group (the amount of storage for this installation may be different). The **Array Management** window should still be displayed after performing the procedure.

The user must be logged on to the SMW as `crayadm` to perform this task.

1. Start the SANtricity Storage Manager.

```
crayadm@smw> /usr/bin/SMclient
```

The SANtricity Storage Manager window appears.

2. If the **Select Addition Method** window appears, choose one of the following options; otherwise, continue with the next step.

- **Automatic** - Select this option if a serial connection was not used to assign IP addresses to the storage array controllers. The SANtricity software automatically detects the available controllers, in-band, using the Fibre Channel link.
- **Manual** - Select this option if IP addresses have already been assigned to the storage array controllers.

NOTE: The following steps assume the **Manual** option was selected.

3. Double-click the name for the storage array that to be configured. The **Array Management** window displays.
4. Select the **Logical/Physical** tab.
5. Right-click **Unconfigured Capacity** and select **Create Volume**. The **Create Volume** wizard displays.
 - a. Select **Next** on the **Introduction (Create Volume)** window.
 - b. Select the **Manual** option on the **Specify Volume Group (Create Volume)** window.
 - c. Select tray 85, slots 1-4 and select **Add**.
 - d. Verify that the RAID level is set to 5.

- e. Select **Calculate Capacity**.
- f. Select **Next** on the **Specify Volume Group (Create Volume)** window.

After creating the first **Volume Group**, create the first volume when prompted.

Create and Configure Volumes for NetApp, Inc. Engenio Devices

Configure the boot RAID with enough LUNs to support the various system management file systems. (Cray recommends a minimum of nine LUNs.)

Three LUNs are required for High Availability SMW. If the CLE feature Direct Attached Lustre (DAL) is configured on SMW HA, another LUN is needed for Image Management and Provisioning System (IMPS) data.

Use the boot LUN configuration table included in *CLE Installation and Configuration Guide (S-2444)*, which is provided with the CLE release package, as a guideline for configuration information for each volume.

NOTE: For a system configured for SMW HA with the SMW failover feature, the boot RAID must also include space for the shared log, MySQL database, and /home file system.

1. Enter a new volume capacity. Specify units as GB or MB.
2. Enter a name for the volume.
3. Select the **Customize Settings** option.
4. Select **Next** in the **Specify Capacity/Name (Create Volume)** window.
5. Verify the settings on the **Customize Advanced Volume Parameters (Create Volume)** window. These settings are used for the all of the LUNs.
 - For **Volume I/O characteristics type**, verify that **File System** is selected.
 - For **Preferred Controller Ownership**, verify that **Slot A** is selected. This places the LUN on the A Controller.
6. Select **Next** in the **Customize Advanced Volume Parameters (Create Volume)** window.
7. In the **Specify Volume to LUN Mapping** window, select the **Default** mapping option.
8. For **Host** type, select **Linux** from the drop-down menu.
9. Select **Finish** in the **Specify Volume to LUN Mapping** window.
10. When prompted to create more LUNs in the **Creation Successful (Create Volume)** window, select **Yes** unless this is the last volume to be created. If this is the last volume, select **No** and skip to step [14](#) on page [62](#)
11. In the **Allocate Capacity (Create Volume)** window, verify that **Free Capacity** is selected on **Volume Group 1 (RAID 5)**.
12. Select **Next** in the **Allocate Capacity (Create Volume)** window.
13. Repeat steps 1 through 13 to create all of the volumes described in the boot LUN configuration table in *CLE Installation and Configuration Guide (S-2444)*, which is provided with the CLE release package.

14. Select **OK** in the **Completed (Create Volume)** window.
15. Create a hot spare. The hot spare provides a ready backup if any of the drives in the Volume Group fail.
 - a. Right-click on the last drive in the slot 14 icon on the right portion of the window and select **Hot Spare Coverage**.
 - b. Select the **Manually Assign Individual Drives** option.
 - c. Select **OK**.
 - d. Select **Close**.
16. Exit the tool.

Configure Remote Logging of NetApp, Inc. Engenio Storage System Boot RAID Messages

The NetApp, Inc. Engenio storage system uses SNMP to provide boot RAID messages. See the NetApp, Inc. Engenio Storage System documentation for additional information.

Zone the QLogic™ FC Switch

For a QLogic Fibre Channel Switch, follow [Configure Zoning for a QLogic SANbox Switch Using QuickTools Utility](#) on page 62 to zone the LUNs on the QLogic SANBox™ switch by using a utility called *QuickTools*.

NOTE: If a LUN is to be shared between failover host pairs, each host must be given access to the LUN. The SMW host port should be given access to all LUNs.

QuickTools is an application that is embedded in the QLogic switch and is accessible from a workstation browser with a compatible Java™ plug-in. It requires a Java browser plugin, version 1.4.2 or later.

These instructions assume that the disk device has four host ports connected to ports 0-3 for the QLogic SANbox switch. The following connections are also required:

- The SMW must be connected to port 10 on the SANBox.
- The boot node must be connected to port 4 on the SANBox.
- The SDB node must be connected to port 5 on the SANBox.
- If a dedicated syslog node is configured, it must be connected to port 6 on the SANBox.

Zoning is implemented by creating a *zone set*, adding one or more zones to the zone set, and selecting the ports to use in the zone.

Prerequisite for this task: the SANBox is configured and on the HSS network.

Configure Zoning for a QLogic SANbox Switch Using QuickTools Utility

1. Start a web browser.
2. Enter the IP address of the switch. If the configuration has a single switch, the IP address is 10.1.0.250. The IP address of each RAID controller is preconfigured by Cray and is listed on a sticker on the back of the RAID controller.
3. Enter the login name and password when the **Add a New Fabric** window pops up and prompts for them. The default administrative login name is `admin`, and the default password is `password`.

4. The QuickTools utility displays in the browser. Select **Add Fabric**.
NOTE: If a dialog box appears stating that the request failed to connect over a secured connection, select **Yes** and continue.
5. The switch is located and displayed in the window. Double-click the switch icon. Information about the switch displays in the right panel.
6. At the bottom of the panel, select the Configured Zonesets tab.
7. From the toolbar menu, select Zoning and then Edit Zoning. The Edit Zoning window displays.
8. Select the Zone Set button. The Create a Zone Set window displays. Create a new zone set. (In this example, assume that the zone set is named XT0.)
9. Right-click the XT0 zone and select Create a Zone.
10. Create a new zone named BOOT.
11. On the right panel, select the button in front of BOOT to open a view of the domain members.
12. Ports 0, 4, 5, and 10 are added to the BOOT zone. Define the ports in the zone to ensure that the discovery of LUNs is consistent among the SMW, the boot node, and the SDB node.
 - a. Using the mouse, left-click Port # 0 and drag it to the BOOT zone.
 - b. Using the mouse, left-click Port # 4 and drag it to the BOOT zone. This port is for the boot node.
 - c. Using the mouse, left-click Port # 5 and drag it to the BOOT zone. This port is for the SDB node.
 - d. Using the mouse, left-click Port # 10 and drag it to the BOOT zone. This port is for the SMW.
13. Select **Apply**. The error-checking window displays.
14. When prompted, select **Perform Error Check**.
15. After confirming that no errors were found, select **Save Zoning**.
16. When prompted to activate a Zone Set, select **Yes** and then select the appropriate XT0 zone set.

At this point, Cray recommends creating a backup of the switch configuration ([Create a Backup of the QLogic Switch Configuration](#) on page 63) before closing and exiting the application.

Create a Backup of the QLogic Switch Configuration

Use the QuickTools utility to create a backup of the QLogic switch configuration. To use QuickTools, a Java browser plugin, version 1.4.2 or later is required.

To start a web browser and open the QuickTools utility, complete steps 1 through 4. If the QuickTools utility is already open, skip to step 5.

1. Start a web browser.
2. Enter the IP address of the switch.
The IP address of each RAID controller is preconfigured by Cray and is listed on a sticker on the back of the RAID controller. If the configuration has a single switch, the IP address is 10.1.0.250.

3. Enter the login name and password when the **Add a New Fabric** window pops up and prompts for them. The RAID default administrative login name is `admin`, and the default password is `password`.
4. The QuickTools utility appears. Select **Add Fabric**. If a dialog box appears stating that the request failed to connect over a secured connection, select **Yes** and continue.
5. From within the QuickTools utility, complete the configuration backup:
 - a. At the top bar, select **Switch** and then **Archive**. A **Save** window pops up with blanks for **Save in:** and **File Name:**.
 - b. Enter the directory (for example, `crayadm`) and a file name (for example, `sanbox_archive`) for saving the QLogic switch configuration.
 - c. Select the **Save** button.
6. Close and exit the application.

Rediscover the LUNs

Reboot the SMW and Verify LUNs are Recognized

This procedure causes the SMW to rediscover the LUNs and zones that were created.

1. Log on as the `root` user.

```
crayadm@smw> su - root
```

2. Reboot the SMW to ensure that the LUNs are recognized:

```
smw# reboot
```



CAUTION: Failure to reboot the SMW at this point could produce unexpected results later on.

3. Log on as the `root` user.

```
crayadm@smw> su - root
```

4. Execute the `lsscsi` command to verify that the LUNs (volumes) have been rediscovered.

```
smw# lsscsi
```

5. List the disk devices by using the `fdisk` command to verify that the LUNs (volumes) are configured according to the boot LUN configuration table in *CLE Installation and Configuration Guide (S-2444)*, which is provided with the CLE release package.

```
smw# fdisk -l
```

Partition the LUNs

After creating, formatting, and zoning the LUNs on the boot RAID, partition them. For procedures and LUN partitioning recommendations, see *CLE Installation and Configuration Guide (S-2444)*, which is provided with the CLE release package.

Determine the Persistent Device Name for a LUN

After initial partitioning of the boot RAID, always address the storage via its persistent `/dev/disk/by-id/` name. Do not use the short `/dev/sdxx` name, which cannot uniquely identify the disk between reboots.

Use this procedure to determine the persistent device name from the LUN number on the boot RAID.

1. Use `lsscsi` to show the `/dev/sd*` device name associated with a LUN number (for example, LUN 15). In the first column of the output, the LUN is the final number in the `[n:n:n:n]` value.

```
crayadm@smw1:~> lsscsi
[0:0:0:0]    disk    ATA      TOSHIBA MK1661GS ME0D  /dev/sda
[0:0:1:0]    disk    ATA      ST91000640NS    AA03  /dev/sdb
[0:0:2:0]    disk    ATA      TOSHIBA MK1661GS ME0D  /dev/sdc
.
.
.
[5:0:0:15]   disk    LSI      INF-01-00      0786  /dev/sdo
[5:0:0:16]   disk    LSI      INF-01-00      0786  /dev/sdp
[5:0:0:17]   disk    LSI      INF-01-00      0786  /dev/sdq
[5:0:0:18]   disk    LSI      INF-01-00      0786  /dev/sdr
```

In this example, LUN 15 is associated with `/dev/sdo`.

2. Use `ls -l` to map the `/dev/sd*` device name to the persistent device name. This example displays the persistent device name for `/dev/sdo` (that is, LUN 15).

```
crayadm@smw1:~> cd /dev/disk/by-id
crayadm@smw1:~> ls -l | grep sdo
lrwxrwxrwx 1 root root 10 Sep  4 00:56 scsi-360080e500037667a000003a2519e3ff2 -
> ../../sdo
lrwxrwxrwx 1 root root 10 Sep  4 00:56 wwn-0x60080e500037667a000003a2519e3ff2 -
> ../../sdo
```

3. Record the LUN numbers for the shared directories so that you can identify the persistent (by-id) device names when you install the CLE software and configure the SMW HA cluster. You may find it helpful to use the following table to record this information.

Table 9. Boot RAID LUNs for the Shared Directories

Directory	LUN Number
<code>/var/lib/mysql</code>	
<code>/var/opt/cray/disk/1</code>	
<code>/home</code>	
<code>/var/lib/pgsql</code> (if on the boot RAID)	

Install CLE Software on the First SMW

The following procedures describe how to install CLE software on the first SMW. For the first SMW, the installation procedures for the CLE software are the same as for a single SMW.

NOTE: If you are converting an existing Cray system (with a single SMW) to an SMW HA cluster, you do not need to reinstall the CLE software. Instead, update the existing SMW to the required CLE release software, then continue to [Install SMW HA Software](#) on page 158.

Prepare to Install a New System

Follow these procedures to perform an initial software installation of the Cray Linux Environment (CLE) 5.2.UP04 software release for a new Cray XC30 system.

Before the CLE Software Installation

Perform the following tasks before you install the CLE 5.2.UP04 software release.

- Review release package documentation. Read the *CLE 5.2.UP04 Release Errata, Limitations for CLE 5.2.UP04* and *README* documents provided with the release for any installation-related requirements and corrections to this installation guide.
- Additional installation information may also be included in *Cray Linux Environment (CLE) Software Release Overview (S-2425)* and *Cray Linux Environment (CLE) Software Release Overview Supplement (S-2497)*.
- Confirm the SMW software release level. You must install the SMW 7.2.UP04 release or later on your SMW before installing the CLE 5.2.UP04 release. If a specific SMW update package is required for your installation, that information is documented in the *README* file provided with the CLE 5.2.UP04 release. The procedures in this guide assume that the SMW software has been successfully installed and the SMW is operational; type the following command to determine the HSS/SMW version:

```
crayadm@smw:~> cat /opt/cray/hss/default/etc/smw-release
7.2.UP04
```

Passwords

The following default account names and passwords are used throughout the CLE software installation process. Cray recommends that you change all default passwords; see [Change the Default System Passwords](#).

Table 10. Default System Passwords

Account Name	Password
root	initial0
crayadm	crayadm

For procedures on handling SMW and RAID accounts and passwords, see *Installing Cray System Management Workstation (SMW) Software (S-2480)*.

Access to MySQL™ databases requires a user name and password. The MySQL accounts and privileges are shown in [MySQL Database Accounts and Privileges](#) on page 67.

Table 11. MySQL Database Accounts and Privileges

Account	Default Password	Privilege
root	None; you must create a password.	All available privileges.
basic	basic	Read access to most tables; most applications use this account.
sys_mgmt	sys_mgmt	Most privileged non-root account; all privileges required to manipulate CLE tables.

For steps to change MySQL account passwords, see [Change Default MySQL Passwords on the SDB](#) on page 116.

Configure the Boot RAID

This chapter describes how to configure, format, zone, and partition the boot RAID (redundant array of independent disks) system.

Cray ships systems with much of this configuration completed. You may not have to perform all of the steps described in this chapter unless you are making changes to the configuration.

Cray provides support for system boot RAID from two different vendors, DataDirect™ Networks (DDN™) and NetApp Corporation. You may also have a QLogic™ SANbox™ Fibre Channel switch from QLogic Corporation or a serial-attached SCSI (SAS) switch from NetApp.

Installing Cray System Management Workstation (SMW) Software (S-2480) contains device specific instructions for configuring boot RAID LUNs (Logical Units) and volume groups.

The DDN RAID uses LUNs; the NetApp, Inc. Engenio™ RAID uses volumes.

If you use NetApp, Inc. Engenio devices for your boot RAID, you must have installed SANtricity™ Storage Manager software from NetApp, Inc. Corporation. For more information about third party software applications required to configure your boot RAID, see *Installing Cray System Management Workstation (SMW) Software (S-2480)*.

After [Configure the Boot RAID LUNs or Volume Groups](#) on page 68, follow the procedures for [Boot LUN Partitions](#) on page 70.

Prerequisites and Assumptions for Configuring the Boot RAID

In typical system installations, the RAID provides the storage for both the boot node root file systems and the shared root file system. Although these file systems are managed from the boot node during normal operation, you must use the SMW to perform an initial installation of the Cray Linux Environment (CLE) base operating system, based on SUSE Linux Enterprise Server (SLES) 11 SP3, and Cray CLE software packages onto the boot RAID disks.

NOTE: For a Cray XC30 system configured for SMW high availability (HA) with the SMW failover feature, the boot RAID is also used for the shared log, MySQL database, and /home file system.

In typical system installations, RAID units provide user and scratch space and can be configured to support a variety of file systems. Different RAID controller models support Fibre Channel (FC), Serial ATA (SATA), and Serial Attached SCSI (SAS) disk options.

The following assumptions are relevant throughout this chapter:

- The SMW has an Ethernet connection to the Hardware Supervisory System (HSS) network.
- The boot node(s) have Ethernet connections to the SMW.
- The SMW has a switched FC or SAS connection to the boot RAID.
- The boot node(s) have a switched FC or SAS connection to the boot RAID.
- The service database (SDB) node(s) have a switched FC or SAS connection to the boot RAID.
- If a dedicated syslog node is configured, it has a switched FC or SAS connection to the boot RAID.

Configure the Boot RAID LUNs or Volume Groups

Follow the procedures in *Installing Cray System Management Workstation (SMW) Software (S-2480)* to configure your boot RAID. You must configure the boot RAID with at least six LUNs; this number ensures that you have enough space to support the various system management file systems and a backup. The recommended configuration listed in [Recommended Boot RAID LUN Values](#) on page 68 describes nine LUNs, spanning three system sets: a backup set for the previous release labeled `BLUE`, a production set for the current release labeled `GREEN`, and a set for testing the deployment of a pending upgrade labeled `RED`. You can specify units as GB or MB.

If you have DDN devices, follow the boot RAID configuration procedures in *Installing Cray System Management Workstation (SMW) Software (S-2480)* to `telnet` to the RAID controller and use the `lun add` and `lun delete` commands to configure LUNs following the recommendations in [Recommended Boot RAID LUN Values](#) on page 68.

If you have NetApp, Inc. Engenio devices, follow the boot RAID configuration procedures in *Installing Cray System Management Workstation (SMW) Software (S-2480)* to use the SANtricity Storage Manager software to create the boot RAID volume group and configure the volumes following the recommendations in [Recommended Boot RAID LUN Values](#).

The example below uses a combined SDB, UFS and `syslog` node; if you are using a dedicated node for either of these functions, they should each have their storage on a separate LUN. A single LUN should not serve multiple physical nodes.

 **WARNING:** Some third-party batch systems require additional space (possibly upwards of 50GB) in the `PERSISTENT_VAR` partition. You should review the requirements of the batch system you intend to deploy in order to determine the appropriate size for this partition and the `shroot*` LUN(s). The size given for the `UFS` partition, and therefore the `sdb*` LUN(s), is based on the assumption that it will be used for the `crayadm` home directory and `/ufs/alps_shared`, and that general users home directories will be on another file system. This is the recommended configuration for best performance. However, if the `UFS` partition will be used by users it should be 40 GB at an absolute minimum and should likely be put on its own separate LUN for better performance.

Table 12. Recommended Boot RAID LUN Values

LUN	Label	Size (1-50 Cabinets)	Size (50+ Cabinets)	Segment Size
0	bootroot0	40GB	70GB	256KB
1	shroot0	280GB	370GB	256KB
2	sdb0	60GB	80GB	256KB
3	bootroot1	40GB	70GB	256KB
4	shroot1	280GB	370GB	256KB

LUN	Label	Size (1-50 Cabinets)	Size (50+ Cabinets)	Segment Size
5	sdb1	60GB	80GB	256KB
6	bootroot2	40GB	70GB	256KB
7	shroot2	280GB	370GB	256KB
8	sdb2	60GB	80GB	256KB

NOTE: For a Cray XC30 system configured for SMW HA with the SMW failover feature, the boot RAID must also include sufficient space for the shared log, MySQL database, and /home file system.

Zone the LUNs

After you configure and format the LUNs, you must grant host access to the LUNs by using a process called *zoning*. Zoning maps a host port on the RAID controller to the LUNs that the host accesses. If you have a QLogic switch, zoning maps the host ports on the switch. Although it is possible to enable all hosts to have access to all LUNs, Cray recommends that each host be granted access only to the LUNs it requires.

NOTE: If a LUN is to be shared between failover host pairs, each host must be given access to the LUN. The SMW host port should be given access to all LUNs.

Zone the LUNs for DDN Devices

If you have DDN devices, follow the procedure to zone LUNs for DDN in *Installing Cray System Management Workstation (SMW) Software (S-2480)*. Use the `zoning` command to edit each port number and map the LUNs; follow the recommendations in [Recommended DDN Zoning](#).

Table 13. Recommended DDN Zoning

Port	External LUN, Internal LUN		
1	000,000	001,001	002,002
2	003,003	004,004	005,005
3	006,006	007,007	008,008
4			

If you have created or modified your LUN configuration, you must reboot the SMW to enable it to recognize the new LUN configuration and zoning information. Verify that all of your changes have been recognized by invoking the `lsscsi` command. [Partition the LUNs](#) on page 71 provides example output for the `lsscsi` command. For more information, see the `lsscsi(8)` man page on the SMW.



WARNING: Failure to reboot the SMW at this point might produce unexpected results. If your SMW does not properly recognize the boot RAID configuration, the system installation procedures could overwrite existing data.

Zone the QLogic FC Switch

If you have a QLogic Fibre Channel Switch, follow the procedures described in *Installing Cray System Management Workstation (SMW) Software (S-2480)* to zone the LUNs on your QLogic SANBox switch. Use the QuickTools utility to create a Zone Set and define the ports in the zone; follow the recommendations in

[Recommended QLogic Zoning](#) on page 70. These recommendations presuppose that the disk device has four host ports connected to ports 0-3 for the QLogic SANbox switch. QuickTools is an application, embedded in your QLogic switch, which is accessible from the SMW by using a web browser.

Zoning for a QLogic switch is implemented by creating a *zoneset*, adding one or more zones to the zone set, and selecting the ports to use in the zone.

Follow this procedure after the SANBox is configured and on the HSS network.

Table 14. Recommended QLogic Zoning

Zone	Port	SANBox Connection
Boot	0	Boot RAID
Boot	4	Boot Node
Boot	5	SDB Node
Boot	10	SMW
Boot	6	Syslog node (if dedicated)

If you have created or modified your LUN configuration, you must reboot the SMW to enable it to recognize the new LUN configuration and zoning information. Verify that all of your changes have been recognized by invoking the `lsscsi` command. [Partition the LUNs](#) on page 71 provides example output for the `lsscsi` command. For more information, see the `lsscsi(8)` man page on the SMW.



WARNING: Failure to reboot the SMW at this point might produce unexpected results. If your SMW does not properly recognize the boot RAID configuration, the system installation procedures could overwrite existing data.

Boot LUN Partitions

After creating, formatting, and zoning the LUNs on the boot RAID, partition them by invoking the `fdisk` command on the SMW.

[Example of Boot LUN Partitions](#) contains an example of a partition layout using the three system sets mentioned above. The SMW Device names in column 5 are consistent with rack-mount SMW hardware with four internal disk drives.



WARNING: Please note that `/dev/sdx` names are not persistent; the names in the example may be different from your system and may change between reboots, so while it is acceptable to use them and the output of `lsscsi` while partitioning, ensure that you are targeting the appropriate device. After initial partitioning, you should always address the storage via its persistent `/dev/disk/by-id/` name. For more information, see [About Persistent Boot RAID Device Names](#) on page 89.

Table 15. Example of Boot LUN Partitions

LUN	System Set	Part Num	Part Type	SMW Device	Size (1-50 Cabinets)	Size (50+ Cabinets)	Type	Description
0	BLUE	1	Primary	sde1	30GB	60GB	Linux	Boot node root file system
0	BLUE	2	Primary	sde2	10GB	10GB	Swap	Boot node swap
1	BLUE	1	Primary	sdf1	210GB	250GB	Linux	Shared root

LUN	System Set	Part Num	Part Type	SMW Device	Size (1-50 Cabinets)	Size (50+ Cabinets)	Type	Description
1	BLUE	2	Primary	sdf2	10GB	10GB	Linux	Boot image 1
1	BLUE	3	Primary	sdf3	10GB	10GB	Linux	Boot image 2
1	BLUE	4	Primary	sdf4	50GB	100GB	Linux	Persistent /var
2	BLUE	1	Primary	sdg1	20GB	20GB	Linux	Service database (sdb)
2	BLUE	2	Primary	sdg2	20GB	20GB	Linux	UFS
2	BLUE	3	Primary	sdg3	20GB	40GB	Linux	syslog
3	GREEN	1	Primary	sdh1	30GB	60GB	Linux	Boot node root file system
3	GREEN	2	Primary	sdh2	10GB	10GB	Swap	Boot node swap
4	GREEN	1	Primary	sdi1	210GB	250GB	Linux	Shared root
4	GREEN	2	Primary	sdi2	10GB	10GB	Linux	Boot image 1
4	GREEN	3	Primary	sdi3	10GB	10GB	Linux	Boot image 2
4	GREEN	4	Primary	sdi4	50GB	100GB	Linux	Persistent /var
5	GREEN	1	Primary	sdj1	20GB	20GB	Linux	Service database (sdb)
5	GREEN	2	Primary	sdj2	20GB	20GB	Linux	UFS
5	GREEN	3	Primary	sdj3	20GB	40GB	Linux	syslog
6	RED	1	Primary	sdk1	30GB	60GB	Linux	Boot node root file system
6	RED	2	Primary	sdk2	10GB	10GB	Swap	Boot node swap
7	RED	1	Primary	sdl1	210GB	250GB	Linux	Shared root
7	RED	2	Primary	sdl2	10GB	10GB	Swap	Boot image 1
7	RED	3	Primary	sdl3	10GB	10GB	Linux	Boot image 2
7	RED	4	Primary	sdl4	50GB	100GB	Linux	Persistent /var
8	RED	1	Primary	sdm1	20GB	20GB	Linux	Service database (sdb)
8	RED	2	Primary	sdm2	20GB	20GB	Linux	UFS
8	RED	3	Primary	sdm3	20GB	40GB	Linux	syslog

Partition the LUNs

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Use the `lsscsi` command to verify that the LUNs were recognized. The first SMW device is the first non-ATA device listed. On an SMW with four internal SATA drives, the output should resemble the following example. Note that four of the disks are ATA, not DDN or NetApp, Inc. Engenio disks. Depending on when the disks were obtained, they might be LSI, Engenio, or NetApp. The `lsscsi` output may be different on your system.

```
smw:~ # lsscsi
[0:0:0:0] disk ATA FUJITSU MHZ2160B 8A22 /dev/
sda
[0:0:1:0] disk ATA ST91000640NS AA02 /dev/sdb
[0:0:2:0] disk ATA FUJITSU MHZ2160B 8A22 /dev/sdc
[0:0:3:0] disk ATA FUJITSU MHZ2160B 8A22 /dev/sdd
[1:0:0:0] disk LSI INF-01-00 0777 /dev/sde
[1:0:0:1] disk LSI INF-01-00 0777 /dev/sdf
[1:0:0:2] disk LSI INF-01-00 0777 /dev/sdg
[1:0:0:3] disk LSI INF-01-00 0777 /dev/sdh
[1:0:0:4] disk LSI INF-01-00 0777 /dev/sdi
[1:0:0:5] disk LSI INF-01-00 0777 /dev/sdj
[1:0:0:6] disk LSI INF-01-00 0777 /dev/sdk
[1:0:0:7] disk LSI INF-01-00 0777 /dev/sdl
[1:0:0:8] disk LSI INF-01-00 0777 /dev/sdm
```

3. Create the partitions shown in [Example of Boot LUN Partitions](#) on page 70 by using the `fdisk` command. If you are not familiar with `fdisk`, see [Configure Primary and Extended File Partitions](#) and the `fdisk(8)` man page.

```
smw:~ # fdisk /dev/sde
```

In this example, repeat the previous command for `/dev/sdf` through `/dev/sdm`; use the values in [Example of Boot LUN Partitions](#) on page 70 for each `fdisk` session. Changes to the partition table are not effective until entering `w` to write and exit.

4. Invoke the `fdisk` command with the `-l` option to verify that the LUNs (volumes) are configured according to [Example of Boot LUN Partitions](#) on page 70. LUN sizes may be slightly different; for example, 43G instead of 40G, as listed in the table. The following output represents the example; output is specific to the actual LUN configuration.

```
smw:~ # fdisk -l
Disk /dev/sda: 160.0 GB, 160041885696 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000081

   Device Boot      Start         End      Blocks   Id  System
/dev/sda1            63      67103504    33551721   82  Linux swap / Solaris
/dev/sda2   *    67103505    312576704    122736600   83  Linux
```

```
Disk /dev/sdb: 1000.2 GB, 1000204886016 bytes
36 heads, 63 sectors/track, 861342 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000083

   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1          2048    1953525167    976761560   83  Linux
```

```
Disk /dev/sdc: 160.0 GB, 160041885696 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000080
```

```
   Device Boot      Start         End      Blocks   Id  System
/dev/sdc1            63      67103504    33551721   82  Linux swap / Solaris
```

```
/dev/sdc2 * 67103505 312576704 122736600 83 Linux
```

```
Disk /dev/sdd: 160.0 GB, 160041885696 bytes
6 heads, 63 sectors/track, 826936 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000082
```

Device	Boot	Start	End	Blocks	Id	System
/dev/sdd1		2048	312581807	156289880	83	Linux

```
Disk /dev/sde: 42.9 GB, 42949672960 bytes
255 heads, 63 sectors/track, 5221 cylinders, total 83886080 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

Disk /dev/sde doesn't contain a valid partition table

```
Disk /dev/sdf: 300.6 GB, 300647710720 bytes
255 heads, 63 sectors/track, 36551 cylinders, total 587202560 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

Disk /dev/sdf doesn't contain a valid partition table

```
Disk /dev/sdg: 64.4 GB, 64424509440 bytes
255 heads, 63 sectors/track, 7832 cylinders, total 125829120 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

Disk /dev/sdg doesn't contain a valid partition table

```
Disk /dev/sdh: 42.9 GB, 42949672960 bytes
255 heads, 63 sectors/track, 5221 cylinders, total 83886080 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

Disk /dev/sdh doesn't contain a valid partition table

```
Disk /dev/sdi: 300.6 GB, 300647710720 bytes
255 heads, 63 sectors/track, 36551 cylinders, total 587202560 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

Disk /dev/sdi doesn't contain a valid partition table

```
Disk /dev/sdj: 64.4 GB, 64424509440 bytes
255 heads, 63 sectors/track, 7832 cylinders, total 125829120 sectors
```

```
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000

Disk /dev/sdj doesn't contain a valid partition table

Disk /dev/sdk: 42.9 GB, 42949672960 bytes
255 heads, 63 sectors/track, 5221 cylinders, total 83886080 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000

Disk /dev/sdk doesn't contain a valid partition table

Disk /dev/sdl: 300.6 GB, 300647710720 bytes
255 heads, 63 sectors/track, 36551 cylinders, total 587202560 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000

Disk /dev/sdl doesn't contain a valid partition table

Disk /dev/sdm: 64.4 GB, 64424509440 bytes
255 heads, 63 sectors/track, 7832 cylinders, total 125829120 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000

Disk /dev/sdm doesn't contain a valid partition table
```

About Installation Configuration Files

This chapter contains essential information about parameters that you must set before you install the Cray Linux Environment (CLE) software on a Cray system. Review this information before installing CLE and again for every CLE software update or upgrade installation.

The CLE software installation process uses an installation script called `CLEinstall`. The `CLEinstall` program, in turn, references two configuration files to determine site-specific configuration parameters used during installation. These configuration files are `CLEinstall.conf` and `/etc/sysset.conf`. Prior to invoking the `CLEinstall` installation program, you must carefully examine these two configuration files and make site-specific changes.



CAUTION: Improper configuration of the `CLEinstall.conf` and `/etc/sysset.conf` files can result in a failed installation.

`CLEinstall.conf`: Based on the settings you define in `CLEinstall.conf`, the `CLEinstall` program updates other configuration files, thus eliminating many manual configuration steps. The `CLEinstall.conf` file is created during the installation process by copying the `CLEinstall.conf` template from the distribution media. This chapter groups the `CLEinstall.conf` settings into three categories: parameters that must be defined for your specific configuration, parameters with default or standard settings that do not need to be changed in most cases, and additional parameters that are required to configure optional functionality or subsystems.

`sysset.conf`: You can install `bootroot` and `sharedroot` to an alternative location while your Cray system is running. This enables you to do the configuration steps in the alternative root location and then move over to the

alternative location after it is configured, thus reducing the need for dedicated system time for installation and configuration. Use the `/etc/sysset.conf` file to identify sets of disk partitions on the boot RAID as alternative *system sets*. Each system set provides a complete collection of all file systems and boot images, thus making it possible to switch easily between two or more versions of the system software. For example, by using system sets, it is possible to keep a stable "production" system available for your users while simultaneously having a "test" system available for new software installation, configuration, and testing.

NOTE: If you have existing `CLEinstall.conf` and `/etc/sysset.conf` files, save copies before you make any changes.

About `CLEinstall.conf` Parameters That Must Be Defined

A template `CLEinstall.conf` is delivered on the Cray CLE 5.2.UP nn Software DVD. Use this sample file to prepare your installation configuration settings before you begin the installation. Carefully examine each installation parameter and the associated comments in the file to determine the changes that are required for your planned configuration.

In [Create the Installation Configuration Files](#) on page 93, you are directed to edit your `CLEinstall.conf` file. Make site-specific changes at that point in the installation process.

These parameters must be changed or verified for your configuration. For more information, see the comments in the `CLEinstall.conf` file or the `CLEinstall.conf(5)` man page.

Mount points on the SMW

Set `bootroot_dir` and `sharedroot_dir` to choose the boot root and shared root file system mount points on the SMW.

Hostname settings

Set `xthostname` and `node_class_login_hostname` to the hostname for your Cray system.

User home directories

Set `home_directory_ufs=no` and enter values for `home_directory_server_hostname`, `home_directory_server_IPaddr`, and `home_directory_server_path` that point to your site-specific user file system NFS™ server; Cray does not recommend using the boot RAID (`/ufs`) for user files on production systems.

IMPORTANT: This section is referred to as "UFS (home directory) for login nodes" in older versions of `CLEinstall.conf`.

Node settings

Set `node_*` parameters to identify which nodes are the `sdb`, `ufs`, `syslog`, `login` and `boot` node(s).

Node class settings

Set `node_class*` parameters to assign nodes to a node class for `/etc/opt/cray/sdb/node_classes`.

NOTE: You must keep the `node_class*` parameters current with the system configuration. Refer to [Maintain Node Class Settings and Hostname Aliases](#) on page 76 for more information.

SSH on boot node settings

Set `ssh_*` parameters to configure boot node root secure shell (`ssh`) keys.

ALPS settings

Set `alps_*` parameters for various Application Level Placement Scheduler (ALPS) configuration options.

GPU Settings

Set `GPU=yes` only if your machine has GPUs installed. Setting this parameter to `yes` will install the required RPMs and code for GPU systems. If your machine has GPU blades without GPUs, or it does not have GPU blades installed, set this parameter to `no`.

For systems with GPUs, you must configure and enable the alternative compute node root run time environment for dynamic shared objects and libraries (DSL). GPU blades require access to shared libraries to support GPUs.

Maintain Node Class Settings and Hostname Aliases

For an initial CLE software installation, the `CLEinstall` program creates the `/etc/opt/cray/sdb/node_classes` file and adds Cray system hostname and alias entries to the `/etc/hosts` file. Additionally, each time you update or upgrade your CLE software, `CLEinstall` verifies the content of `/etc/opt/cray/sdb/node_classes` and modifies `/etc/hosts` to match the configuration specified in your `CLEinstall.conf` file.

Unless you confirm that your hardware changed, the `CLEinstall` program fails if `/etc/opt/cray/sdb/node_classes` does not agree with `node_classidx` parameters in `CLEinstall.conf`. Therefore, you must keep the following parameters current with your Cray system configuration:

```
node_class_login=login
```

Specifies the node class label for the login nodes.

```
node_class_default=service
```

Specifies the default node class label for service nodes. A service node can only be in one class; typical classes might be `service`, `login`, `network`, `sdb`, `ost`, `mds`, or `lustre`. Classes can have any name provided the names are used consistently by using the `xtspec` command. Node IDs that are not designated as part of a class default to `node_class_default`.

```
node_class[idx]=class NID NID ...
```

Specifies the name of the class for index `idx` and the integer node IDs (NIDs) that belong to the class. `CLEinstall` uses `node_class[idx]` parameters along with other parameters in `CLEinstall.conf` to create, update or verify `/etc/opt/cray/sdb/node_classes` and `/etc/hosts` files. You must configure a `login` class with at least one NID. A NID can be a member of only one `node_class`.

`CLEinstall` uses the information you specify for these parameters to update the `/etc/hosts` file as follows:

- A copy of the original file is saved as `/etc/hosts.$$preinstall`.
- The Cray system entries (IP address, node ID, and physical name) are moved to the end of the file.
- Any Cray hostname aliases specified in `CLEinstall.conf` are added for the appropriate nodes.
- A copy of the modified file is saved as `/etc/hosts.$$postinstall`.

Set the `node_class[idx]` parameters

```
node_class[0]=login 8 30
node_class[1]=network 9 13 27 143
node_class[2]=sdb 5
node_class[3]=lustre 12 18 26
```

For each class defined, host name aliases in `/etc/hosts` are assigned based on the class name and order of NIDs specified for this parameter.

Host alias assignments based on the `node_class[idx]` parameters

If you define the following `node_class` class entry:

```
node_class[1]=network 9 13 27 143 19
```

Host name aliases for the `network` class are assigned as follows:

```
nid00009 - network1
nid00013 - network2
nid00027 - network3
nid00143 - network4
nid00019 - network5
```

About `CLEinstall.conf` Parameters with Standard Settings

The standard or default values for settings in the following categories are appropriate in many cases. Verify that these default values are acceptable for your site. For more information, see the comments in the `CLEinstall.conf` file or the `CLEinstall.conf(5)` man page.

- Shared root setting
- Boot node network settings
- SDB node network settings
- Persistent `var` settings
- `syslog` settings
- Partition setting
- SDB database settings
- Writeable `/tmp` for CNL setting
- Writeable `/var/tmp` for CNL setting
- Node Health Check (NHC) on boot
- NHC communication over Secure Sockets Layer (SSL)

Additional parameters that you should review are described in greater detail in the following sections.

Change the Default High-speed Network (HSN) Settings

By default, the HSN IP address is `10.128.0.0`. You can modify these parameters to configure another valid address; for example `10.33.0.0`. In most cases, the default value is acceptable. Modify the following HSN settings as needed:

```
HSN_byte1=
HSN_byte2=
```

The `HSN_byte1=` and `HSN_byte2=` parameters specify the HSN IP address. The default values are `HSN_byte1=10`, `HSN_byte2=128`. Cray recommends that the values for `HSN_byte1` and `HSN_byte2` do not overlap subnets listed as default IP addresses in *Installing Cray System Management Workstation (SMW) Software (S-2480)*.

When `HSN_byte1` and `HSN_byte2` are changed from the default, CLEinstall implements this change by modifying the following files:

- `/etc/sysconfig/xt` on the boot root and shared root
- `/etc/hosts` on the boot root and shared root
- `/etc/sysconfig/alps` on the boot root and shared root
- `/etc/opt/cray/rca/fomd.conf` on the boot root and shared root
- `/etc/opt/cray/hosts/service_alias.conf` on the boot root and shared root
- `/opt/xt-images/templates/default/etc/hosts` for CNL and SNL images
- `/opt/xt-images/templates/default/etc/krsip.conf` for CNL images with RSIP

In addition, the CNL parameters file and the SNL parameters file in the bootimage are updated to include `bootnodeip`, `sdbnodeip`, `ippob1` and `ippob2`.



CAUTION: Because of site-specific local modifications, additional files might require updating when the HSN IP address changes. For example, such files as `/etc/hosts.allow`, `/etc/hosts.deny`, `/etc/exports`, and `/etc/security/access.conf` might require updating.

`bootimage_bootifnetmask`

This netmask must be consistent with the modified `HSN_byte1` and `HSN_byte2` parameters.

```
persistent_var_IPaddr
home_directory_server_IPaddr
bootnode_failover_IPaddr
bootimage_bootnodeip
alps_directory_server_IPaddr
```

The `HSN_byte1` and `HSN_byte2` parameters and the netmask must be consistent with the first two bytes of these IP addresses that are defined in `CLEinstall.conf`.

Change `home_directory_server_IPaddr` only if `home_directory_ufs=no`; change `alps_directory_server_IPaddr` only if `alps_directory_server_hostname` is not the `ufs` node hostname.

Change Parameters to Tune Virtual Memory or NFS

You may choose to modify these parameters based on your system configuration.

`sysctl_conf_vm_min_free_kbytes`

Specifies the `vm_min_free_kbytes` parameter of the Linux kernel. Linux virtual memory must keep a minimum number of kilobytes free. The virtual memory uses this number to compute a `pages_min` value for each `lowmem` zone in the system. Based on this value, each `lowmem` zone is allocated a number of reserved free pages, in proportion to its size.

The default value of `vm_min_free_kbytes` in the `/etc/sysctl.conf` file is 102,400 KB of free memory. For some configurations, the default value may be too low, and memory exhaustion may occur even though free memory is available. If this happens, adjust the `vm_min_free_kbytes` parameter to increase the value to 5% or 6% of total memory.

`nfs_mountd_num_threads`

Controls an NFS `mountd` tuning parameter that is added to `/etc/sysconfig/nfs` and used by `/etc/init.d/nfsserver` to configure the number of `mountd` threads on the boot node. By default, NFS `mountd` behavior is unchanged (a single thread). For systems with more than 50 service I/O nodes, Cray recommends that you configure multiple threads by setting this parameter to 4. If you have a larger Cray system (greater than 50 service I/O nodes), contact your Cray service representative for assistance changing the default setting.

use_kernel_nfsd_number

Specifies the number of NFSD threads. By default, this variable in `/etc/sysconfig/nfs` is set to 16.

A large site may wish to change both `nfs_mountd_num_threads` and `use_kernel_nfsd_number`. Contact your Cray service representative for assistance changing the default setting.

nfsserver

Specifies whether or not the `nfsserver` service should be enabled on all service nodes. If this parameter is set to `yes`, all service nodes will run the `nfsserver` service. If this parameter is set to `no`, only the boot, `sdb`, and `ufs` nodes will run the `nfsserver` service. The default value is `no`.

Change the Default bootimage Settings

You can change several parameters related to the boot image configuration. In most cases, the default values are acceptable. For information about additional bootimage parameters, see the `CLEinstall.conf(5)` man page.

bootimage_temp_directory=/home/crayadm/boot

Specifies the parent directory on the SMW for temporary directories used to extract a boot image and adjust the boot image parameters file.

bootimage_bootnodeip=10.131.255.254

Specifies the virtual IP address for the boot node. The default is `10.131.255.254`. In most cases, the default value is acceptable. If you change the default, you must also modify default value for `bootnode_failover_IPaddr` and `persistent_var_IPaddr` to match the address specified by `bootimage_bootnodeip`.

bootimage_bootifnetmask=255.252.0.0

Specifies the network mask for the boot node virtual IP address. The default is `255.252.0.0`. In most cases, the default value is acceptable.

bootimage_xtrel

Set this parameter to `yes` to add `xtrel=$XTrelease` to the boot image SNL parameters file. This option is used for release switching; for more information see the `xtrelswitch(8)` man page. The default value is `no`.

Change Turbo Boost Limit

Because processors have a high degree of variability in the amount of turbo boost each processor can supply, limiting the amount of turbo boost can reduce performance variability and reduce power consumption. The limit applies only when a high number of cores are active. On an N-core processor, the limit is in effect when the active

core count is N, N-1, N-2, or N-3. On a 12-core processor, the limit is in effect when 12, 11, 10, or 9 cores are active.

NOTE: Turbo boost is not supported on SandyBridge processors.

Set the following CLEinstall.conf parameter to adjust the turbo boost limit.

```
turbo_boost_limit=999
```

The valid values are 100, 200 and 999. The default setting is 999. When `turbo_boost_limit=100`, 100 MHz is the limit. A value of 200 limits turbo boost to 200 MHz. A value of 999 implies no turbo boost limit is applied.

Node Health on Boot

Node Health Checker (NHC) automatically checks the health of compute nodes on boot using the Node Health Checker. The `NHC_on_boot` variable controls this feature and is set to `NHC_on_boot=yes` by default.

The `NHC_on_boot` variable affects the NHC configuration file in the compute node image and is used only when NHC is run on boot. Every NHC invocation after the compute node boot, either by ALPS or manually, uses the configuration file on the shared root.

If your site does not have a site customized file

in `/opt/xt-images/templates/default/etc/opt/cray/nodehealth/nodehealth.conf` for node health on boot, then the a sample one is copied into place there. You should modify the file in the template directory for your site.

Enable NHC Communication Over Secure Sockets Layer (SSL)

The `NHC_SSL` parameter is set to `yes` by default in `CLEinstall.conf`. This parameter generates the appropriate keys for Node Health Checker (NHC) to use SSL for communication with the compute nodes. Cray recommends sites configure NHC to use SSL.

The key (`rsa_key`), certificate (`rsa_cert`), and the certificate signing request (`servercsr`) are created in the `/root/.nodehealth` directory on the shared root and copied to the `/opt/xt-images/templates/default-px/root/.nodehealth` directory for the compute node images. If any one of these three required files are missing from the shared root, they are all generated again and copied back out. If the `NHC_SSL` setting in `CLEinstall.conf` is set to `no`, the files are removed from both locations.

About CLEinstall.conf Parameters for Additional Features and Subsystems

The `CLEinstall.conf` file contains settings for optional functionality and subsystems. To configure and enable a particular functionality, its settings require modifications. Settings for unused features and subsystems can be ignored.

For more information, see the comments in the `CLEinstall.conf` file or the `CLEinstall.conf(5)` man page.

Lustre File System Support and Tuning

The Lustre file system is optional; however, applications that run on CNL compute nodes require either Lustre file systems or DVS in order to perform I/O operations. Several `CLEinstall.conf` parameters are available to configure your system for Lustre file systems and set up basic Lustre file system tuning. In most cases, the default values are acceptable. In addition to setting these parameters, refer to [Install and Configure Direct-Attached Lustre](#), as you complete the installation or upgrade process.

```
lustre_elevator=noop
```

Specifies a value for `elevator` in the SNL boot image parameters file; sets the default scheduler for a Lustre object storage server (OSS). Currently, the `noop` scheduler is recommended for Lustre on high-performance storage.

```
lustre_clients=
```

Specifies a value for `max_nodes` in `/etc/modprobe.conf.local` for service nodes; used to calculate buffer allocation for connection to Lustre clients. Cray recommends setting this parameter to the total number of compute nodes and login nodes configured on the Cray system, rounded up to the nearest 100.

```
lustre_servers=
```

Specifies a value for `max_nodes` in `/opt/xt-images/templates/default/etc/modprobe.conf` for compute nodes; used to calculate buffer allocation for connection to Lustre servers. Cray recommends that you set this parameter to the total number of Lustre servers configured on your Cray system, rounded up to the nearest 100.

```
lustre_credits=2048
```

Specifies a value for `credits` in `/etc/modprobe.conf.local` for service nodes; defines the number of outstanding transactions allowed for a Lustre server. Cray recommends that you set this parameter to 2048.

```
lustre_peer_hash_table_size=509
```

Specifies a value for `peer_hash_table_size` in `/etc/modprobe.conf.local` for service nodes; defines the size of the hash table for the client peers and enables `lnet` to search large numbers of peers more efficiently. Cray recommends that you set this parameter to 509.

```
lustre_oss_num_threads=256
```

Specifies a value for `lustre_oss_num_threads` in `/etc/modprobe.conf.local` for service nodes; defines the number of threads a Lustre OSS uses. Cray recommends that you set this parameter to 256 threads.

```
direct_attached_lustre=no
```

When this setting is `yes`, then direct-attached Lustre (DAL) is enabled. `CLEinstall` uses IMPS to prepare the image to be added to the unified bootimage so that nodes chosen to be internal Lustre service nodes can be booted with the CentOS operating system with Cray additions for Lustre servers. When this setting is `yes`, the `--Centosmedia` option is required when running `CLEinstall`. `CLEinstall` provides command hints for the rest of the DAL configuration steps.

Configure Boot Node Failover

Boot-node Failover is an optional CLE feature that sets up a backup boot node to automatically take over when the primary boot node fails.

Set these parameters to configure `CLEinstall` to automatically complete several configuration steps for boot-node failover.

In addition, specify the primary and backup nodes in the boot configuration and configure the STONITH capability on the blade or module of the primary boot node. These tasks are done after creating boot images later in the new install, update, or upgrade processes.

The following `CLEinstall.conf` parameters configure boot-node failover.

```
node_boot_alterate=
```

Specifies the backup or alternate boot node. The alternate boot node requires an Ethernet connection to the SMW and a QLogic Host Bus Adapter (HBA) card to communicate with the boot RAID. The alternate boot node must not reside on the same blade as the primary boot node.

```
bootnode_failover=yes
```

Set this parameter to `yes` to configure boot-node failover.



CAUTION: The STONITH capability is required to implement boot-node failover. Because STONITH is a per blade setting and not a per node setting, you must ensure that your primary boot node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

```
bootnode_failover_IPaddr=10.131.255.254
bootimage_bootnodeip=10.131.255.254
persistent_var_IPaddr=10.131.255.254
```

Specifies the virtual IP address for boot-node failover. These must all match. The default is `10.131.255.254`. In most cases, the default value is acceptable. You must modify the default value for the other two parameters to match the address specified by `bootnode_failover_IPaddr`.

```
bootnode_failover_netmask=255.252.0.0
```

Specifies the network mask for the boot-node failover virtual IP address. The default is `255.252.0.0`. In most cases, the default value is acceptable.

```
bootnode_failover_interface=ipogif0:1
```

Specifies the virtual network interface for boot-node failover. The default value is `ipogif0:1`. In most cases, the default value is acceptable.

For additional information, including manual boot node failover configuration steps, see *Managing System Software for the Cray Linux Environment* (S-2393).

Configure SDB Node Failover

Service database (SDB) Node failover is an optional CLE feature that enables automatic failover to a backup SDB node when the primary SDB node fails.

Use the parameters described in this section to configure `CLEinstall` to automatically complete several configuration steps for SDB node failover.

When these parameters are used to configure SDB node failover, the `CLEinstall` program will verify and turn on `chkconfig` services and associated configuration files for `sdbfailover`.

The backup SDB node uses the `/etc` files that are class or node specialized for the primary SDB node and not for the backup node itself; the `/etc` files for the backup node are identical to those that existed on the primary SDB node.

For additional information about SDB node failover, see *Managing System Software for the Cray Linux Environment* (S-2393).

In addition, configure STONITH for the primary SDB node, specify the primary and backup nodes in the boot configuration, and optionally create a site-specific `sdbfailover.conf` file for the backup SDB node. These tasks are done after creating boot images later in the new install, update, or upgrade processes.

After booting and testing your system, follow [Configure Boot Automation for SDB Node Failover](#) on page 140 to configure your system to start SDB services automatically on the backup SDB node in the event of a SDB node failover.

The following `CLEinstall.conf` parameters configure SDB node failover.

```
node_sdb_alternate=
```

Specifies the backup or alternate SDB node. The alternate SDB node requires a QLogic Host Bus Adapter (HBA) card to communicate with the RAID. This node is dedicated and cannot be used for other service I/O functions. The alternate SDB node must reside on a separate blade from the primary SDB node.

```
sdbnode_failover=yes
```

Set this parameter to `yes` to configure SDB node failover.



CAUTION: The STONITH capability is required to implement SDB node failover. Because STONITH is a per blade setting and not a per node setting, you must ensure that your primary SDB node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

```
sdbnode_failover_IPaddr=10.131.255.253
```

Specifies the virtual IP address for SDB node failover. The default is `10.131.255.253`. In most cases, the default value is acceptable.

```
sdbnode_failover_netmask=255.252.0.0
```

Specifies the network mask for the SDB node failover virtual IP address. The default is `255.252.0.0`. In most cases, the default value is acceptable.

```
sdbnode_failover_interface=ipogif0:1
```

Specifies the virtual network interface for SDB node failover. This parameter must be defined even if you are not configuring SDB node failover. The default value is `ipogif0:1`. In most cases, the default value is acceptable.

Include DVS in the Compute Node Boot Image

The following `CLEinstall.conf` parameter configures `CLEinstall` to include the Data Virtualization Service (DVS) RPM in the compute node boot image. Cray DVS is an optional CLE feature. In addition to setting this parameter, refer to [Configure Cray DVS](#) on page 128, as you complete the installation or upgrade process.

```
CNL_dvs=yes
```

Set this parameter to `yes` to include the DVS RPM in the compute node boot image.

Optionally, edit the `/var/opt/cray/install/shell_bootimage_LABEL.sh` script and specify `CNL_DVS=y` before updating the CNL boot image.

For additional information about DVS, see [Introduction to Cray Data Virtualization Service \(S-0005\)](#).

Configure DSL and CNRTE

When the CLE compute node root runtime environment (CNRTE) is configured, users can link and load dynamic shared objects in their applications. To configure and install the compute node root runtime environment, configure the shared root as a DVS-projected file system. Dynamic shared objects and libraries (DSL) and the compute node root runtime environment (CNRTE) are optional.

To configure DSL and the compute node root runtime environment for your Cray system, follow these steps.

1. Select the service or compute nodes to configure as compute node root servers. Any compute nodes used for CNRTE will no longer be part of the compute node pool. Do not use the same nodes configured as Lustre server nodes.
2. Modify DSL-specific parameters according to the system configuration by editing the `CLEinstall.conf` file ([Create the Installation Configuration Files](#) on page 93).
3. Configuring compute nodes as compute node root servers requires additional configuration. Cray recommends configuring the nodes as repurposed compute nodes. Complete [Repurpose Compute Nodes as Service Nodes](#) on page 94 before running `CLEinstall`.

The `CLEinstall` program creates a default `cnos` specialization class. This class allows an administrator to specialize files specifically for compute nodes; it is used with dynamic shared objects and libraries (DSL). If the `cnos` specialization class exists and DSL is enabled, those specialized `/etc` files are automatically mounted on the compute node roots.

For additional information about DSL, see *Managing System Software for the Cray Linux Environment* (S-2393). For additional information about DVS, see *Introduction to Cray Data Virtualization Service* (S-0005).

Set the following parameters in the `CLEinstall.conf` file to cause the `CLEinstall` program to automatically configure the system for the compute node root runtime environment.

`DSL=yes`

Set this parameter to `yes` to enable dynamic shared objects and libraries and the CNRTE. The default is `no`. Setting this option to `yes` enables DVS.

`DSL_nodes=`

Specifies the nodes that will act as DVS compute node root servers. These nodes can be a combination of service or compute nodes. Set to integer node IDs (NIDs) separated by a space.

`DSL_mountpoint=/dsl`

Specifies the mount point on the DVS servers for the compute nodes; it is the projection of the shared root file system. The compute nodes mount this path as `/`. In most cases, the default value is acceptable.

`DSL_attrcache_timeout=14400`

Specifies the attribute cache time out for compute node root servers; it is the number of seconds before DVS attributes are considered invalid and are retrieved from the server again. In most cases, the default value is acceptable.

Configure Realm-Specific IP Addressing (RSIP)

Realm-Specific IP Addressing allows CLE compute and service nodes to share IP addresses configured on the external Gigabit and 10 Gigabit Ethernet interfaces of network nodes. RSIP is an optional CLE feature. By sharing the external addresses, you may rely on your system's use of private address space and avoid the need to

configure compute nodes with addresses within your site's IP address space. The external hosts see only the external IP addresses of the Cray system.

RSIP on Cray systems supports IPv4 TCP and UDP transport protocols but not IP Security and IPv6 protocols.

Select the nodes to configure as RSIP servers. RSIP servers must run on service nodes that have a local external IP interface such as a 10GbE network interface card (NIC). Cray requires that you configure RSIP servers as dedicated network nodes.



WARNING: Do not run RSIP servers on service nodes that provide Lustre services, login services, or batch services.

The following `CLEinstall.conf` parameters configure RSIP.

```
rsip_nodes=
```

Specifies the RSIP servers. Populate with space separated integer NIDs of the nodes you have identified as RSIP servers.

```
rsip_interfaces=
```

Specifies the IP interface for each RSIP server node. Populate with a space separated list of interfaces that correlate with the `rsip_nodes` parameter.

```
rsip_servicenode_clients=
```

Set this parameter to a space separated integer list of service nodes to use for RSIP clients.



WARNING: Do not configure service nodes with external network connections as RSIP clients. Configuring a network node as an RSIP client will disrupt network functionality. Service nodes with external network connections will route all non-local traffic into the RSIP tunnel and IP may not function as desired.

```
CNL_rsip=yes
```

Set this parameter to `yes` to include the RSIP RPM in the compute node boot image. Optionally, you can edit the `/var/opt/cray/install/shell_bootimage_LABEL.sh` script and specify `CNL_RSIP=y` before you update the CNL boot image.

For example, to configure `nid00016` and `nid00020` as RSIP servers both using an external interface named `eth0`; `nid00064` as an RSIP server using an external interface named `eth1`; and `nid00000` as a service node RSIP client, set the following parameters.

```
rsip_nodes=16 20 64
rsip_interfaces=eth0 eth0 eth1
rsip_servicenode_clients=0
CNL_rsip=yes
```

For additional information, see the `rsipd(8)`, `xtrsipcfg(8)`, and `rsipd.conf(5)` man pages and *Managing System Software for the Cray Linux Environment (S-2393)*. Enhancements to the default RSIP configuration require a detailed analysis of site-specific configuration requirements. Contact your Cray representative for assistance in changing the default RSIP configuration.

Configure Service Node MAMU

Service Node Multiple Application Multiple User (MAMU) support provides the ability to set aside a small number of re-purposed compute nodes for serial workload. Serial workload nodes can be configured in advance, while the node is booted as a service node. See *Managing System Software for the Cray Linux Environment* for more details on this feature.



CAUTION: If you use the CLE installer, the nodes you specify as service MAMU nodes must be compute nodes when you start their configuration and installation.

The following `CLEinstall.conf` parameters configure service node MAMU.

```
SERVICE_MAMU=no
```

Set this parameter to `yes` to use a set of nodes to serve as a separate workload management pool of execution (MOM) nodes. These execution nodes do not run ALPS or manage Cray Workload, but are available for core level-scheduling directly through the workload manager.

```
SERVICE_MAMU_classes=postproc
```

Uncomment this line to set up a default class, `postproc`, for serial workload nodes. To specify more than one class, use a space to separate the class names. For example, `SERVICE_MAMU_classes=postproc serial`.

```
node_class[6]=postproc 15 16 17 18
```

Uncomment this line, which appears in the node class settings section of `CLEinstall.conf`.

Define service node MAMU NIDs

Use this class entry to set up nodes with NIDs 15, 20, 33, and 37.

```
node_class[n]=postproc 15 20 33 37
```

The value of `n` is the next unused node class defined in `CLEinstall.conf`. Configure a node class for each class listed in the `SERVICE_MAMU_classes` setting.

Configure Compute Node Swap

```
CNL_swap=no
```

Set `CNL_swap=yes` to enable the configuration of compute nodes as service nodes. Enabling the swap parameter causes `CLEinstall` to install two RPMs, one on the compute node image and the other on the shared root for the service nodes.

Configure Graphics Processing Units

The following `CLEinstall.conf` parameters configure graphics processing units (GPUs).

```
GPU=no
```

Set `GPU=yes` only if your machine has GPUs installed. Setting this parameter to `yes` will install the required RPMs and code for GPU systems. If your machine has GPU blades without GPUs, or it does not have GPU blades installed, set this parameter to `no`.

For systems with GPUs, you must configure and enable the alternative compute node root run time environment for dynamic shared objects and libraries (DSL). GPU blades require access to shared libraries to support GPUs. The parameters are `DSL=yes`, `DSL_nodes`, `DSL_mountpoint`, and `DSL_attrcache_timeout`. For more information, see [Configure DSL and CNRTE](#) on page 83.

Configure Intel Xeon Phi Coprocessors

The following `CLEinstall.conf` parameters configure Intel Xeon Phi™ coprocessors.

KNC=no

Specifies whether Xeon Phi coprocessors are present in the system. Setting `KNC=yes` enables support for Xeon Phi compute blades. The default value is `no`.

KNC_BASE=50000

Defines an offset that is used to prepare hostnames and IP addresses for the Xeon Phi nodes. If the CNL node hosting a Xeon Phi coprocessor is `nid00032`, then the coprocessor has a hostname of `nid50032` and a hostname alias of `acc50032`.

Also notice that the `cname` for the Intel MIC has `a0` added to the `c0-0c0s8n0` hostname. Below are the two entries in `/etc/hosts` for a compute node and the Xeon Phi coprocessor node hosted by that compute node.

```
10.128.0.33      nid00032      c0-0c0s8n0
10.128.196.249  nid50032      c0-0c0s8n0a0  acc00032
```

Configure `ntpclient` for Clock Synchronization

A network time protocol (NTP) client, `ntpclient`, is available to install on compute nodes; it synchronizes the time of day on the compute node clock with the clock on the boot node. The `ntpclient` is an optional CLE feature.

Without this feature, compute node clocks drift apart over time. When `ntpclient` is installed, the clocks drift apart during a four hour calibration period and then converge on the time reported by the boot node. Note that the standard CLE configuration includes an NTP daemon (`ntpd`) on the boot node to synchronize with the clock on the SMW, and the service nodes run `ntpd` to synchronize with the boot node.

Use the following `CLEinstall.conf` parameter to enable `ntpclient` on the compute nodes.

CNL_ntpclient=yes

Set this parameter to `yes` to include the `ntpclient` RPMs in the compute node boot image.

Optionally, you can edit the `/var/opt/cray/install/shell_bootimage_LABEL.sh` script and specify `CNL_NTPCLIENT=y` before you update the CNL boot image.

Configure the Parallel Command (`pcmd`) Tool for Unprivileged Users

Parallel Command (`pcmd`) is a secure tool that runs commands on the compute nodes as the user who launched the command. A user can specify which nodes to run the command on. Configuring the Parallel Command Tool for unprivileged users is an optional CLE feature. Sites that are uncomfortable having a `setuid` root program on their system may keep `pcmd` a root-only tool. For more information, see the `pcmd(1)` man page.

NHC_pcmd_suid=no

The `pcmd` is installed as a root-only tool by default. To allow non-root users to run the tool, `pcmd` can be installed as a `setuid` root program. Do this at installation time by specifying `NHC_pcmd_suid=yes` in the `CLEinstall.conf` file.

Configure High Speed Network Metrics

NOTE: Configuring network metrics monitoring is an optional feature for Cray XC30, Cray XE, and Cray XK systems. Also, using the Lightweight Distributed Metric Service (LDMS) within OVIS to collect network metrics is optional. Administrators can configure network metric collection without using OVIS as a client application. However, this feature is provided with LDMS metric aggregation in mind and in that use case LDMS is responsible for aggregating and collecting compute node metrics to the SMW. For more information on installing and configuring LDMS and OVIS on Cray systems, see <https://ovis.ca.sandia.gov/mediawiki/index.php/CRAY-LDMS>.

High speed network (HSN) metrics monitoring is a CLE feature that provides on-node metric collection and aggregation for system nodes. Cray provides kernel modules and utilities for metric collection. Per-NIC HSN metrics collected include: injection and ejection bandwidths, kernel output and input bandwidths. For each ASIC link, dimension metrics collected include: reception data, packet counts (XE/XK only), time stalled, and lane status. The values provided are not continuous values but by using associated timestamps, rates can be determined.

Use the following `CLEinstall.conf` parameter to enable network metric collection.

```
CNL_network_metrics=no
```

Set `CNL_network_metrics` to `yes` in `CLEinstall.conf` to enable network metric collection. For more information on network metric collection, see *Managing System Software for the Cray Linux Environment*.

About System Set Configuration in `/etc/sysset.conf`

The `/etc/sysset.conf` configuration file defines system sets. Each system set is defined by the following information for each device or boot RAID disk partition in the set: *function*, *SMWdevice*, *host*, *hostdevice*, *mountpoint*, and a *shared* flag. Each system set definition also contains a `LABEL` and a `DESCRIPTION`. The information regarding the disk partition is based on the zoning of the LUNs on the boot RAID.

A system administrator can use this file to configure a group of disk devices and disk partitions on the boot RAID into a system set that can be used as a complete bootable system. By configuring system sets, a system administrator can easily switch between different software releases or configurations. For example, you can use (or create) separate production and test system sets to manage updates and upgrades of the CLE operating system.

In [Create Configuration Files](#) on page 92, you are directed to create a `/etc/sysset.conf` file specifically for your system configuration. A sample or template file for `/etc/sysset.conf` is delivered on the Cray CLE 5.2.UPnn Software DVD (also see [/etc/sysset.conf Examples](#) for reference). The template contains two example system sets (`BLUE` and `GREEN`). Modify these examples to match your system configuration. You must create the `/etc/sysset.conf` file before you invoke the installation program, at which time you specify the system set to install, upgrade, or update.

Follow these requirements, restrictions, and tips when you create a site-specific `sysset.conf` file. For more information, see the `sysset.conf(5)` man page.

- The `/etc/sysset.conf` file includes two sets of device names for the boot RAID; *SMWdevice* is the pathname to the disk partition on the SMW and *hostdevice* is the pathname on the Cray system (host).
- You must configure persistent device names for the boot RAID disk devices. Cray recommends that you use the `/dev/disk/by-id/` persistent device names (or LVM device names for LVM devices, which are also

persistent). For more information, see [About Persistent Boot RAID Device Names](#) on page 89. For more information about LVM configuration, see [Configure LVM for System Backups](#).

- Some partitions may be shared between two or more sets, such as `/syslog`.
- Some partitions must exist in only one set; for example, a matched triplet of boot root, shared root, and boot image.
- `SMWdevice` may be a path name to a device or a dash (-).
- `hostdevice` may be a path name to a device or a dash (-).
- Set `SMWdevice` and `hostdevice` to dash (-) for `BOOT_IMAGEn` if the boot image is a file and not a raw device.
- `hostdevice` may be a dash (-) with a real `SMWdevice` only when the `function` is `RESERVED`.
- `BOOT_IMAGEn` may be a raw disk device that has `SMWdevice` and `hostdevice` as path names to real devices. Specify `mountpoint` as a link to that device.
- `BOOT_IMAGEn` may be an archive (`cpio`) file in a directory. The directory must exist on both the SMW and the boot root, with the same name. Specify `mountpoint` as the path name to this type of boot image file.
- `mountpoint` may be a dash (-) if it is a Lustre device (`LUSTREMDS0` or `LUSTREOST0`).
- The `RESERVED` `function` can be used to indicate that a partition has a site-defined function and should not be overwritten by `CLEinstall` or `xthotbackup`.
- Some partitions may be marked `RESERVED` and yet belong to a system set.
- The system set `LABEL` contains all orphaned disk partitions that are not in any other system set.
- If the SMW does not have access to the `SDB` and `SYSLOG` disk devices on the boot RAID, specify `SMWdevice` for these entries as a dash (-). Ensure `hostdevice` is set to the node that has access to these disk partitions. In this case, the `CLEinstall` program generates scripts to create these file systems and suggests when to run the scripts.

In [Create the Installation Configuration Files](#) on page 93, you are directed to create and edit your `/etc/sysset.conf` file. Make all site-specific changes at that point in the installation process.

About Device Partitions in `/etc/sysset.conf`

Check the boot RAID configuration and QLogic switch zoning (for QLogic Fibre Channel switch or DDN device), SAS switch zoning, or SANshare configuration (for NetApp, Inc. disks). These can be configured to allow all hosts to see all LUNs or to allow some hosts to see only a few LUNs.

Use the `fdisk` command on the SMW to confirm that your partitions are identified. Invoke `fdisk -l` to display a list of all detected partitions on the boot RAID disk devices. Compare the output to the list of `SMWdevice` partitions included in the `/etc/sysset.conf` file. Identify any partitions without an assigned `function` and confirm that they are unused. You may include these remaining partitions in the system set labeled `RESERVED` in `/etc/sysset.conf`.

About Persistent Boot RAID Device Names

The `/etc/sysset.conf` file includes two sets of device names for the boot RAID; `SMWdevice` and `hostdevice`. Because SCSI device names (`/dev/sd*`) are not guaranteed to be numbered the same from boot to boot, you must configure persistent device names for these boot RAID disk devices. Cray recommends that you use the `/dev/disk/by-id` persistent device names or LVM device names (if your system is configured to use LVM), which are also persistent.



CAUTION: You must use `/dev/disk/by-id` or an LVM device name when specifying the root file system. There is no support in the `initramfs` for `cray-scsidev-emulation` or custom `udev` rules.

To configure persistent `by-id` device names, modify the `SMWdevice` and `hostdevice` columns to match the `/dev/disk/by-id/` SCSI device names on your system.

The code that follows is the system set format from the `sysset.conf` template:

```
# LABEL:
# DESCRIPTION:
# function      SMWdevice      host      hostdevice      mountpoint      shared
# BOOTNODE_ROOT /dev/disk/by-id/IDa-part1 boot /dev/disk/by-id/IDa-part1 / no
# BOOTNODE_SWAP /dev/disk/by-id/IDa-part2 boot /dev/disk/by-id/IDa-part2 swap no
# SHAREDROOT    /dev/disk/by-id/IDc-part6 boot /dev/disk/by-id/IDc-part6 /xr no
# BOOT_IMAGE0   /dev/disk/by-id/IDc-part7 boot /dev/disk/by-id/IDc-part7 /raw0 no
# BOOT_IMAGE1   - boot - /bootimagedir/xt.tst1 no
# BOOT_IMAGE2   - boot - /bootimagedir/xt.tst2 no
# BOOT_IMAGE3   - boot - /bootimagedir/xt.tst3 no
# SDB           /dev/disk/by-id/IDd-part1 sdb /dev/disk/by-id/IDd-part1 /var/lib/mysql no
# SYSLOG        /dev/disk/by-id/IDe-part1 syslog /dev/disk/by-id/IDe-part1 /syslog no
# UFS           /dev/disk/by-id/IDf-part2 ufs /dev/disk/by-id/IDf-part2 /ufs no
# PERSISTENT_VAR /dev/disk/by-id/IDc-part9 boot /dev/disk/by-id/IDc-part9 /snv no
# LUSTREMSD0    /dev/disk/by-id/IDc-part5 nid00008 /dev/disk/by-id/IDc-part5 - no
# LUSTREOST0    /dev/disk/by-id/IDh-part1 nid00011 /dev/disk/by-id/IDh-part1 - no
```

Modifying `/etc/sysset.conf` for persistent `by-id` device names

When you create a site-specific `/etc/sysset.conf` file ([Create the Installation Configuration Files](#) on page 93), modify each device path to use the persistent device names in `/dev/disk/by-id`.

For each partition identified in [About Device Partitions in `/etc/sysset.conf`](#) on page 89, determine the `by-id` persistent device name. For example, if you defined the boot node root and swap to be devices `sdc1` and `sdc2`, invoke the following commands and note the volume identifier portion of the names.

```
crayadm@smw:~> ls -l /dev/disk/by-id/* | grep sdc
lrwxrwxrwx 1 root root 9 2010-02-23 14:35 /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21 -> ../../sdc
lrwxrwxrwx 1 root root 10 2010-02-23 14:35 /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part1 -> ../../sdc1
lrwxrwxrwx 1 root root 10 2010-02-23 14:35 /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part2 -> ../../sdc2
crayadm@smw:~>
```

Replace `IDa-part*` for both `SMWdevice` and `hostdevice` with the volume identifier and partition number. For example, change:

```
# BOOTNODE_ROOT /dev/disk/by-id/IDa-part1 boot /dev/disk/by-id/IDa-part1 / no
# BOOTNODE_SWAP /dev/disk/by-id/IDa-part2 boot /dev/disk/by-id/IDa-part2 swap no
```

to

```
BOOTNODE_ROOT /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part1 boot \
/dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part1 / no
BOOTNODE_SWAP /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part2 boot \
/dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part2 swap no
```

Ensure that each entry is on a single line. For formatting purposes, the example splits each entry into two lines.

Modified system set with persistent device names

```
LABEL:MYCRAYPRD
DESCRIPTION: mycray production system set
# function      SMWdevice      host \
                hostdevice      mountpoint      shared
BOOTNODE_ROOT  /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part1 boot \
                /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part1 / no
BOOTNODE_SWAP  /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part2 boot \
                /dev/disk/by-id/scsi-3600a0b800026e1400000192a4b66eb21-part2 swap no
SHAREDROOT     /dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part1 boot \
```

BOOT_IMAGE0	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part1	/rr	no
	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part2	boot \	
	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part2	/raw0	no
BOOT_IMAGE1	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part3	boot \	
	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part3	/raw1	no
PERSISTENT_VAR	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part4	boot \	
	/dev/disk/by-id/scsi-3600a0b800026e1400000192c4b66eb70-part4	/snv	no
SDB	/dev/disk/by-id/scsi-3600a0b800026e1400000192e4b66eb97-part1	sdb \	
	/dev/disk/by-id/scsi-3600a0b800026e1400000192e4b66eb97-part1	/var/lib/mysql	no
UFS	/dev/disk/by-id/scsi-3600a0b800026e1400000192e4b66eb97-part2	ufs \	
	/dev/disk/by-id/scsi-3600a0b800026e1400000192e4b66eb97-part2	/ufs	no
SYSLOG	/dev/disk/by-id/scsi-3600a0b800026e140000019304b66ebbb-part1	syslog \	
	/dev/disk/by-id/scsi-3600a0b800026e140000019304b66ebbb-part1	/syslog	no

Ensure that each entry is on a single line. For formatting purposes, the example splits each entry into two lines.

Install CLE on a New System

This chapter contains the information and procedures that are required to perform an initial installation of the Cray Linux Environment (CLE) base operating system (based on SLES 11 SP3) and Cray CLE software packages on a new Cray system.

After you have configured, formatted, zoned, and partitioned the RAID, follow the steps in this chapter to install the system software on the boot RAID partitions. Perform this work on the SMW.



WARNING: The procedures in this chapter install the operating system software on your Cray system. You will overwrite existing CLE system software on the SMW and on the designated system partitions. If you are already running CLE software on your system, see [Prepare to Update or Upgrade CLE Software](#).

Install CLE Software on the SMW

Three DVDs are required to install the CLE 5.2.UP04 release on a Cray system. The first is labeled `Cray CLE 5.2.UPnn Software` and contains software specific to Cray systems. Optionally, you may have an ISO image called `xc-sles11sp3-n.n.nnavv.iso`, where `n.n.nn` indicates the CLE release build level, and `avv` indicates the installer version.

The second DVD is labeled `Cray-CLEbase11sp3-yyyymmdd` and contains the CLE 5.2 base operating system which is based on SLES 11 SP3. The third DVD is labeled `CentOS-6.5-x86_64-bin-DVD1.iso` and contains the CentOS 6.5 base operating system for CLE direct-attached Lustre (DAL) nodes.

Copy the Software to the SMW

1. If the Cray system is booted, use your site-specific procedures to shut down the system. For example, to shutdown using an automation file:

```
crayadm@smw:~> xtbootstys -s last -a auto.xtshutdown
```

For more information about using automation files, see the `xtbootstys(8)` man page.

Although not the preferred method, alternatively execute these commands as `root` from the boot node to shutdown your system.

```
boot:~ # xtshutdown -y
boot:~ # shutdown -h now;exit
```

2. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

3. Insert the Cray CLE 5.2.UPnn Software DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the release media using the ISO image, execute the following command, where `xc-sles11sp3-5.2.55d05.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro xc-sles11sp3-5.2.55d05.iso /media/cdrom
```

4. Copy all files to a directory on the SMW in `/home/crayadm/install.xtre1`, where `xtrel` is a site-determined name specific to the release being installed.

```
smw:~# mkdir /home/crayadm/install.5.2.55
smw:~# cp -pr /media/cdrom/* /home/crayadm/install.5.2.55
```

5. Unmount the Cray CLE 5.2.UPnn Software DVD and eject it.

```
smw:~# umount /media/cdrom
smw:~# eject
```

6. Insert the Cray-CLEbase11sp3 DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the base operating system media using the ISO image, execute the following command, where `Cray-CLEbase11sp3-yyyymmdd.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro Cray-CLEbase11sp3-yyyymmdd.iso /media/cdrom
```

Install CLE software on the SMW

1. As `root`, execute the `CRAYCLEinstall.sh` script to install the Cray CLE software on the SMW.

```
smw:~# /home/crayadm/install.5.2.55/CRAYCLEinstall.sh \
-m /home/crayadm/install.5.2.55 -v -i -w
```

2. At the prompt 'Do you wish to continue?', type `y` and press Enter.

The output of the installation script displays on the console. If this script fails, restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. Do not be concerned about these messages.

NOTE: If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

Create Configuration Files



CAUTION: [About Installation Configuration Files](#) on page 74 contains essential information about specific parameters that you must set before you install CLE software on a Cray system. Read it carefully before continuing. Improper configuration of the `CLEinstall.conf` and `/etc/sysset.conf` files can result in a failed installation.

As noted in [About System Set Configuration in /etc/sysset.conf](#) on page 88, the CLE 5.2 release software can be installed on a system that has never had the CLE 5.2 release installed on it, or the release can be installed to an alternative root location.

If this is the first installation, creating the `CLEinstall.conf` and `/etc/sysset.conf` configuration files is required. After the first installation is complete, any installations to the alternative root location can use the `/etc/sysset.conf` file that was created during the first installation.

When installing direct-attached Lustre, make sure that the `CLEinstall.conf` specifies this setting:
`direct_attached_lustre=yes.`

To configure a system for future LVM backups, follow the procedures in [Configure LVM for System Backups](#) before the CLE install. Install CLE on the LVM-configured system set and you'll be able to use LVM snapshots for future system backups.

Based on the settings you choose in the `CLEinstall.conf` file, the `CLEinstall` program updates other configuration files. The `/etc/sysset.conf` file describes the assignment of devices and disk partitions on the boot RAID and their file systems or functions. For a description of the contents of these files, see [About Installation Configuration Files](#) on page 74 or the `sysset.conf(5)` and `CLEinstall.conf(5)` man pages.

Log out and back in again to access man pages that were installed in [Install CLE software on the SMW](#) on page 92.

Create the Installation Configuration Files

1. Edit the `/home/crayadm/install.xtre1/CLEinstall.conf` configuration file. Carefully follow [About Installation Configuration Files](#) on page 74 and make modifications for your specific configuration.

```
smw:~# chmod 644 /home/crayadm/install.5.2.55/CLEinstall.conf
smw:~# vi /home/crayadm/install.5.2.55/CLEinstall.conf
```

TIP: Use the `rtr --system-map` command to translate between node IDs (NIDs) and physical ID names.

2. Copy the `/home/crayadm/install.xtre1/sysset.conf` system set template file to `/etc/sysset.conf`.



CAUTION: If there is already an `/etc/sysset.conf` file from a previous installation or upgrade, skip this step and do not overwrite it.

```
smw:~# cp -p /home/crayadm/install.5.2.55/sysset.conf /etc/sysset.conf
```

3. Edit the `/etc/sysset.conf` file so that it describes the disk devices and disk partitions that have been previously created on the boot RAID; designate the function or file system for each disk device and disk partition.

```
smw:~# chmod 644 /etc/sysset.conf
smw:~# vi /etc/sysset.conf
```



CAUTION: You must ensure that *SMWdevice* and *hostdevice* are configured with persistent device names, based on your configuration. For more information, see [About Persistent Boot RAID Device Names](#) on page 89 and the `sysset.conf(5)` man page.

- a. For each function, determine the persistent `by-id` device names for your system by using the following command. For a complete example, see [About Persistent Boot RAID Device Names](#) on page 89.

```
crayadm@smw:~> ls -l /dev/disk/by-id/* | grep sdc
```

- b. Modify the `SMWdevice` and `hostdevice` columns to match the `/dev/disk/by-id/` SCSI device names on your system.
4. Make all site-specific changes; for example, configure separate production and test system sets. Save the file. For more information, see [About System Set Configuration in /etc/sysset.conf](#) on page 88.

Repurpose Compute Nodes as Service Nodes

CLE and SMW software include functionality to optionally change the role of compute nodes and boot the hardware with service node images. Use this functionality to add service nodes for services that do not require external connectivity, such as `DSL_nodes`. When a compute node is configured with a service node role, that node is referred to as a *repurposed compute node*.

Do not repurpose compute nodes that are intended to be service MAMU nodes until after running the `CLEinstall` program. For more information, see [Configure Service Node MAMU](#) on page 85.

The Cray system hardware state data is maintained in an HSS database where each node is marked with a compute or service node role. By using the `xtcli mark_node` command, you can mark a node in a compute blade to have a role of `service`.

Because they are marked as service nodes within the HSS, repurposed compute nodes are initialized as service nodes by the `CLEinstall` program and are booted automatically when all service nodes are booted.

Mark Repurposed Compute Nodes as Service Nodes in the HSS

Repeat the following steps for each NID you want to repurpose, for example, compute nodes as `DSL_nodes`.

1. Mark the repurposed compute node as a service node by using the `xtcli mark_node` command. For example:

```
crayadm@smw:~> xtcli mark_node service c0-0c0s7n0
```

2. Verify that the node is a service node by using the `xtcli status` command. For example:

```
crayadm@smw:~> xtcli status c0-0c0s7n0
Network topology: class 0
Network type: Aries
      Nodeid: Service  Core Arch|  Comp state      [Flags]
-----
      c0-0c0s7n0: service MC24  OP|           on      [noflags|]
-----
crayadm@smw:~>
```

Run the CLEinstall Program

The CLEinstall program installs and performs basic configuration of the CLE software for your configuration by using information in the CLEinstall.conf and sysset.conf configuration files.

The CLEinstall program accepts the following options:

`--label=system_set_label`

This option is required. Specify the label of the system set to be used for this installation. The specified label must exist in the system set configuration file that is specified with the `--syssetfile` option. This label is case-sensitive.

`--install | --upgrade | --bootimage-only | --reconfigure`

This option is required. For full installations, use the `--install` option. For upgrade or update installations, use the `--upgrade` option. The `--upgrade` option requires that specifying the release with `--XTrrelease=release_number` and Cray recommends that you also use the `--CLEmedia` option to specify a release-specific directory for the CLE software media. The `--bootimage-only` option recreates the `shell_bootimage_LABEL.sh` script and performs no other installation or upgrade related tasks. For a reconfiguration of CLE features or hardware changes, use the `--reconfigure` option. This option requires that you specify the release with `--XTrrelease=release_number`, and Cray recommends that you also use the `--CLEmedia` option to specify a release-specific directory for the CLE software media.

`--syssetfile=system_set_configuration_file`

Specify the system set configuration file. The default is `/etc/sysset.conf`.

`--configfile=CLEinstall_configuration_file`

Specify the installation configuration file. The default is `./CLEinstall.conf`.

`--nodebug`

Turn off debugging output to a debug file. By default, debugging output is written to `/var/adm/cray/logs/CLEinstall.debug.timestamp`.

`--Basemedia=directory`

Specify which directory the CLE base operating system media is mounted on. The default is `/media/cdrom`.

`--CLEmedia=directory`

Specify the directory where the software media has been placed. The default is `/home/crayadm/install`. The `--CLEmedia` option is required if the media is not in the default location. Documented installation procedures place the software media in a release-specific directory; for example, `/home/crayadm/install.release_number`, therefore, Cray recommends that you always use this option.

`--XTrrelease=release_number`

Specify the CLE release and build level. *release_number* is a string in the form `x.y.level`, where *level* is the unique build identifier; for example, 5.2.55.

NOTE: The `--XTrrelease` option is required with the `--upgrade` and `--reconfigure` options, and it is not valid with the `--install` option.

`--xthwinxmlfile=XT_hardware_inventory_XML_file`

Specify the hardware inventory XML file to use in place of the output from the `xthwinv` command with the `-x` option.

By default, CLEinstall invokes the `xthwinv -x` command on the SMW to retrieve hardware component information and creates a file, `/etc/opt/cray/sdb/attr.xthwinv.xml`, on the boot root file system. When this option is specified, *XT_hardware_inventory_XML_file* is copied to `/etc/opt/cray/sdb/attr.xthwinv.xml` and the

`xthwinv -x` command is not invoked. The `/etc/opt/cray/sdb/attr.xthwinv.xml` file is used in conjunction with the `/etc/opt/cray/sdb/attr.defaults` file to populate the node attributes table of the Service Database (SDB).

Use this option when the Cray system hardware is unavailable, or when you configure a backup SMW that is not connected to the Hardware Supervisory System (HSS) network, or when you configure an unavailable partition on a partitioned system.

The `XT_hardware_inventory_XML_file` must contain output from the `xthwinv -x` command.

`--bootparameters=file`

Specify the service node boot parameters file to be used when making the service node boot image. The `CLEinstall` program modifies this file as needed to include parameters that are defined in the `CLEinstall.conf` or `sysset.conf` configuration files. If this option is not specified for a new system installation, the default parameters file is used. If this option is not specified for an upgrade installation, the parameters file within an existing boot image is used.

`--CNLbootparameters=file`

Specify the CNL compute node boot parameters file to be used when making the CNL boot image. The `CLEinstall` program modifies this file as needed to include parameters that are defined in the `CLEinstall.conf` or `sysset.conf` configuration files. If this option is not specified for a new system installation, the default parameters file is used. If this option is not specified for an upgrade installation, the parameters file within an existing boot image is used.

`--Lustreversion=version_number`

Specify the version of Lustre to be installed on the CLE nodes running the Lustre client. For example, 2.4. If this option is not specified, `CLEinstall` will use the default version of Lustre. The version specified here does not affect the version of Lustre server that runs on direct-attached Lustre nodes.

`--Centosmedia=directory`

Specify the directory where the CentOS™ software media has been mounted. The `--Centosmedia` option is required when installing or upgrading CLE with direct-attached Lustre. For example, the CentOS image mount point could be `/media/Centosbase`.

`--noforcefsck`

Prevent `CLEinstall` from forcing a file system check. If this option is specified, `CLEinstall` invokes the `fsck` command without the `-f` option. This option is not recommended for normal use. Specify this option when restarting `CLEinstall` after resolving an error.

`--version`

Display the version of the `CLEinstall` program.

`--help`

Display help message.

This information is also available in the `CLEinstall(8)` man page.

Run CLEinstall

1. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` ISO image if it is not already mounted.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Also include the `--Centosmedia=directory` option when invoking `CLEinstall`. In this example, the option is `--Centosmedia=/media/Centosbase`.

2. Invoke the `CLEinstall` program on the SMW. `CLEinstall` is located in the directory you created in [Copy the Software to the SMW](#) on page 91.

```
smw:~# /home/crayadm/install.5.2.55/CLEinstall --install --label=system_set_label \
--configfile=/home/crayadm/install.5.2.55/CLEinstall.conf \
--CLEmedia=/home/crayadm/install.5.2.55
```

3. Examine the initial messages directed to standard output. Log files are created in `/var/adm/cray/logs` and named by using a timestamp that indicates when the install script began executing. For example:

```
08:57:48 Installation output will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.stdout.log
08:57:48 Installation errors (stderr) will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.stderr.log
08:57:48 Installation debugging messages will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.debug.log
```

The naming conventions of these logs are:

`CLEinstall.p#.YYYYMMDDhhmmss.$LABEL.logtype.log`

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format.

`$LABEL` is the system set label (in the example above, `CLE52-P3`).

`logtype` is `stdout` (standard output), `stderr` (standard error), or `debug`.

Also, log files are created in `/var/adm/cray/logs` each time `CLEinstall` calls `CRAYCLEinstall.sh`. For example:

```
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.01-B.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.02-B.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.03-B.log
.
.
.
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.17-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.18-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.19-S.log
```

The naming conventions of these logs are:

`CRAYCLEinstall.sh.p#.YYYYMMDDhhmmss.$LABEL.sequence#-root.log`

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format. This is the same timestamp used for the log files of the `CLEinstall` program instance that called `CRAYCLEinstall.sh`.

`$LABEL` is the system set label.

`sequence#` is an increasing count that specifies each invocation of `CRAYCLEinstall.sh` by `CLEinstall`.

`root` is either `B` (bootroot) or `S` (sharedroot), specifying the root modified by the `CRAYCLEinstall.sh` call.

4. CLEinstall validates `sysset.conf` and `CLEinstall.conf` configuration settings and then confirms the expected status of your boot node and file systems.

Confirm that the installation is proceeding as expected, respond to warnings and prompts, and resolve any issues. For example:

- If you are installing to a system set that is not running, and you did not shut down your Cray system, respond to the following warning and prompt:

```
WARNING: At least one blade of p0 seems to be booted.
Please confirm that the system set you are intending
to update is not booted.
Do you wish to proceed?[n]:
```



WARNING: If the boot node has a file system mounted and CLEinstall on the SMW creates a new file system on that disk partition, the running system will be corrupted.

- If you have configured file systems that are shared between two system sets, respond to the following prompt to confirm creation of new file systems:

```
09:21:24 INFO: The PERSISTENT_VAR disk function for the LABEL system set is marked shared.
09:21:24 INFO: The /dev/sdr1 disk partition will be mounted on the SMW for PERSISTENT_VAR
disk function. Confirm that it is not mounted on any nodes in a running XT
system before continuing.
Do you wish to proceed?[n]:y
```

- If the `node_class_idx` parameters do not match the existing `/etc/opt/cray/sdb/node_classes` file, CLEinstall will abort and require you to correct the node class configuration in `CLEinstall.conf` and/or the `node_classes` file on the `bootroot` and/or `sharedroot`. Correct the file, unmount the file systems and rerun CLEinstall:

```
09:21:41 WARNING: valid service node 56 of class server_dvs from
/bootroot0/etc/opt/cray/sdb/node_classes \
is not in CLEinstall.conf and is not the default class service.
09:21:41 INFO: There is one WARNING about a discrepancy between CLEinstall.conf
and /bootroot0/etc/opt/cray/sdb/node_classes.
09:21:41 FATAL: Correct the node class settings discrepancy between CLEinstall.conf
and /bootroot0/etc/opt/cray/sdb/node_classes and restart CLEinstall
```

CLEinstall may resolve some issues after you indicate that you want to proceed; for example, disk devices are already mounted, boot image file or links already exist, HSS daemons are stopped on the SMW.



CAUTION: Some problems can be resolved only through manual intervention via another terminal window or by rebooting the SMW; for example, a process is using a mounted disk partition, preventing CLEinstall from unmounting the partition.

5. Monitor the debug output. Create another terminal window and invoke the `tail` command by using the path and timestamp displayed when CLEinstall was run.

```
smw~:# tail -f /var/adm/cray/logs/CLEinstall.p#. YYYYMMDDhhmmss. $LABEL.debug.log
```

6. Locate the following warning and prompt in the CLEinstall console window and type `y`.

```
*** Preparing to INSTALL software on system set label system_set_label. This will
DESTROY any existing data on disk partitions in this system set. Do you wish to
proceed? [n]
```

7. The CLEinstall program now installs the release software. This process takes a long time; CLEinstall runs from 30 minutes to 1½ hours, depending on your specific system configuration. Monitor the output to ensure that your installation is proceeding without error.

NOTE: Several error messages from the tar command are displayed as the persistent `/var` is updated for each service node. You may safely ignore these messages.

8. Confirm that the CLEinstall program has completed successfully.

On completion, the CLEinstall program generates a list of suggested commands to be run as the next steps in the installation process. These commands are customized, based on the variables in the `CLEinstall.conf` and `sysset.conf` files, and include runtime variables such as PID numbers in file names. The list of suggested commands is written to the `/var/adm/cray/logs/CLEinstall.P#.YYYYMMDDhhmmss.$LABEL.command_hints.log` file in the installer log directory.

Complete the installation and configuration of your Cray system by using both the commands that the CLEinstall program provides and the information in the remaining sections of this chapter and in [Install and Configure Direct-Attached Lustre](#).

As you complete these procedures, you can cut and paste the suggested commands from the output window or from the window created in a previous step that tailed the debug file. The log files created in `/var/adm/cray/logs` for `CLEinstall.P#.YYYYMMDDhhmmss.$LABEL.stdout.log` and `CLEinstall.P#.YYYYMMDDhhmmss.$LABEL.debug.log` also contain the suggested commands.

Create Boot Images

The Cray CNL compute nodes and Cray service nodes use RAM disks for booting. Service nodes and CNL compute nodes use the same `initramfs` format and workspace environment. This space is created in `/opt/xt-images/machine-xtrelease-LABEL-partition/nodetype`, where `machine` is the Cray hostname, `xtrelease` is the build level for the CLE release, `LABEL` is the system set label used from `/etc/sysset.conf`, `partition` describes either the full machine or a system partition, and `nodetype` is either `compute` or `service`.



CAUTION: Existing files in `/opt/xt-images/templates/default` are copied into the new bootimage work space. In most cases, you can use the older version of the files with the upgraded system. However, some file content may have changed with the new release. Verify that site-specific modifications are compatible. For example, use existing copies of `/etc/hosts`, `/etc/passwd` and `/etc/modprobe.conf`, but if `/init` changed for the template, the site-modified version that is copied and used for CLE 5.2 may cause a boot failure.

Follow the procedures in this section to prepare the work space in `/opt/xt-images`. For more information about configuring boot images for service and compute nodes, see the `xtclone(8)` and `xtpackage(8)` man pages.

Modify Boot Image Parameters for Service Nodes

The CLEinstall program modifies a `parameters` file for service nodes located in the `bootimage_temp_directory`.

If the `bootimage_temp_directory` is `/home/crayadm/boot`, the modified `parameters` file is:

```
/home/crayadm/boot/bootimage.default.LABEL.xtrelease.timestamp/SNL0/parameters
```

For example, the default `parameters` file is:

```
/home/crayadm/boot/bootimage.default.CLE52.5.2.14.201403060906/SNL0/parameters-snl
```

The contents of the `parameters` file is a single line, but the following example is formatted here for readability.

```
earlyprintk=ttyS0,115200
load_ramdisk=1
```

```

ramdisk_size=80000
console=ttyS0,115200n8
bootnodeip=10.131.255.254
bootproto=ipog
bootpath=/rr/current
rootfs=nfs-shared
root=/dev/disk/by-id/scsi-3600a0b800051215e000003a84b4ad820-part1
pci=lastbus=3
oops=panic
bootifnetmask=255.252.0.0
elevator=noop
ippob1=10
ippob2=128
iommu=off
pci=noacpi
bad_page=panic
sdbnodeip=10.131.255.253

```

1. Inspect the modified parameters file. In most cases, this file does not need to be changed.

```
smw:~# cat /home/crayadm/boot/bootimage.default.21822/SNL0/parameters
```

2. If you need to change one or more of the variables that are not set from `CLEinstall.conf` or `sysset.conf`, edit the parameters file.

```
smw:~# vi /home/crayadm/boot/bootimage.default.21822/SNL0/parameters
```

Prepare Compute and Service Node Boot Images

Invoke the `shell_bootimage_LABEL.sh` script to prepare boot images for the system set with the specified `LABEL`. When `shell_bootimage_LABEL.sh` is run, it creates a log file in `/var/adm/cray/logs/shell_bootimage_LABEL.sh.PID.log`. This script uses `xtclone` and `xtpackage` to prepare the work space in `/opt/xt-images`.

`shell_bootimage_LABEL.sh` accepts the following options:

- v** Run in verbose mode.
- h** Display help message.
- c** Create and set the boot image for the next boot. The default is to display `xtbootimg` and `xtcli` commands that will generate the boot image. Use the `-c` option to invoke these commands automatically.
- b *bootimage*** Specify *bootimage* as the boot image disk device or file name. The default *bootimage* is determined by using values for the system set `LABEL` when `CLEinstall` was run. Use this option to override the default and manage multiple boot images.
- C *coldstart_dir*** Specify *coldstart_dir* as the path to the HSS coldstart applets directory. The default is `/opt/hss-coldstart+gemini/default/xt` for Cray XE systems. Use this option to override the default. This option is not applicable to Cray XC30 systems. For more information, see the `xtbounce(8)` man page.

Optionally, this script includes `CNL_*` parameters that can be used to modify the CNL boot image configuration defined in `CLEinstall.conf`. Edit the script and set the associated parameter to `y` to load an optional RPM or change the `/tmp` configuration.

1. Run `shell_bootimage_LABEL.sh`, where `LABEL` is the system set label specified in `/etc/sysset.conf` for this boot image. For example, if the system set label is `BLUE`, log on to the SMW as `root` and type:

```
smw:~# /var/opt/cray/install/shell_bootimage_BLUE.sh
```

On completion, the script displays the `xtbootimg` and `xtcli` commands that are required to build and set the boot image for the next boot. If the `-c` option was specified, the script invokes these commands automatically and the remaining steps in this procedure should be skipped.

2. Create a unified boot image for compute and service nodes by using the `xtbootimg` command suggested by the `shell_bootimage_LABEL.sh` script.

In the following example, replace `bootimage` with the `mountpoint` for `BOOT_IMAGE0` in the system set that is defined in `/etc/sysset.conf`. Set `bootimage` to either a raw device; for example `/raw0` or a file name; for example `/bootimagedir/bootimage.new`.



CAUTION: If `bootimage` is a file, verify that the file exists in the same path on both the SMW and the boot root.

For Cray XC30 systems, type this command (for partitioned systems, replace `s0` with `pN` where `N` is the partition number for which the image is being built):

```
smw:~# xtbootimg \
-L /opt/xt-images/hostname-xtrelease-LABEL-s0/compute/CNL0.load \
-L /opt/xt-images/hostname-xtrelease-LABEL-s0/service/SNL0.load \
-c bootimage
```

- a. At the prompt 'Do you want to overwrite', type `y` to overwrite the existing boot image file.
- b. If `bootimage` is a file, copy the boot image file from the SMW to the same directory on the boot root. If `bootimage` is a raw device, skip this step. For example, if the `bootimage` file is `/bootimagedir/bootimage.new` and `bootroot_dir` is set to `/bootroot0`, type the following command:

```
smw:~# cp -p /bootimagedir/bootimage.new /bootroot0/bootimagedir/bootimage.new
```

3. Set the boot image for the next system boot by using the suggested `xtcli` command.

The `shell_bootimage_LABEL.sh` program suggests an `xtcli` command to set the boot image based on the value of `BOOT_IMAGE0` for the system set being used. The `-i bootimage` option specifies the path to the boot image and is either a raw device; for example, `/raw0` or `/raw1`, or a file such as `/bootimagedir/bootimage.new`.



CAUTION: The next boot, anywhere on the system, uses the boot image set here.

- a. Display the currently active boot image. Record the output of this command.

If the partition variable in `CLEinstall.conf` is `s0`, type:

```
smw:~# xtcli boot_cfg show
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type:

```
smw:~# xtcli part_cfg show pN
```

- b. Invoke `xtcli` with the `update` option to set the default boot configuration used by the boot manager.

If the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the boot image to be used for the entire system.

```
smw:~# xtcli boot_cfg update -i bootimage
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command to select the boot image to be used for the designated partition.

```
smw:~# xtcli part_cfg update pN -i bootimage
```

Enable Boot Node Failover

NOTE: Boot node failover is an optional CLE feature.

If boot-node failover has been configured for the first time, follow these steps. If boot-node failover has not been configured, skip this procedure.

To enable bootnode failover, you must set `bootnode_failover` parameters in the `CLEinstall.conf` file before you run the `CLEinstall` program. For more information, see [Configure Boot Node Failover](#) on page 81.

In this example, the primary boot node is `c0-0c0s0n1` (`node_boot_primary=1`) and the backup or alternate boot node is `c0-0c1s1n1` (`node_boot_alternate=61`).

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

1. As `crayadm` on the SMW, halt the primary and alternate boot nodes.



WARNING: Verify that the system is shut down before you invoke the `xtcli halt` command.

```
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
```

2. Specify the primary and backup boot nodes in the boot configuration.

If the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the boot node for the entire system.

```
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command to select the boot node for the designated partition.

```
crayadm@smw:~> xtcli part_cfg update pN -b c0-0c0s0n1,c0-0c1s1n1
```

3. To use boot-node failover, enable the STONITH capability on the blade or module of the primary boot node. Use the `xtdaemonconfig` command to determine the current STONITH setting.

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary boot node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

4. To enable STONITH on the primary boot node blade, type the following command:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 stonith=true
c0-0c0s0: stonith=true
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.

Enable SDB Node Failover

NOTE: SDB node failover is an optional CLE feature.

If SDB node failover has been configured for the first time, follow these steps. If SDB node failover has not been configured, skip this procedure.

In addition to this procedure, refer to [Configure Boot Automation for SDB Node Failover](#) on page 140 after you have completed the remaining configuration steps and have booted and tested your system.

To enable SDB node failover, you must set `sdbnode_failover` parameters in the `CLEinstall.conf` file before you run the `CLEinstall` program. For more information, see [Configure SDB Node Failover](#) on page 82.

In this example, the primary SDB node is `c0-0c0s2n1` (`node_sdb_primary=5`) and the backup or alternate SDB node is `c0-0c1s3n1` (`node_sdb_alternate=57`).

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

1. Invoke `xtdaemonconfig` to determine the current STONITH setting on the blade or module of the primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary SDB node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

2. Enable STONITH on your primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s2 stonith=true
c0-0c0s2: stonith=true
The expected response was received.
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.

3. Specify the primary and backup SDB nodes in the boot configuration.

For example, if the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the primary and backup SDB nodes.

```
crayadm@smw:~> xtcli halt c0-0c0s2n1,c0-0c1s3n1
crayadm@smw:~> xtcli boot_cfg update -d c0-0c0s2n1,c0-0c1s3n1
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command:

```
crayadm@smw:~> xtcli part_cfg update pN -d c0-0c0s2n1,c0-0c1s3n1
```

Run Post-CLEinstall Commands

1. Unmount and eject the release software DVD from the SMW DVD drive if it is still loaded.

```
smw:~# umount /media/cdrom
smw:~# umount /media/Centosbase
smw:~# eject
```

2. Run the `shell_post_install.sh` script on the SMW to unmount the boot root and shared root file systems and perform other cleanup as needed.

```
smw:~# /var/opt/cray/install/shell_post_install.sh /bootroot0 /sharedroot0
```



WARNING: Exercise care when you mount and unmount file systems. If you mount a file system on the SMW and boot node simultaneously, you may corrupt the file system.

3. Confirm that the `shell_post_install.sh` script successfully unmounted the boot root and shared root file systems.

If a file system does not unmount successfully, the script displays information about open files and associated processes (by using the `lsdf` and `fuser` commands). Attempt to terminate processes with open files and if necessary, reboot the SMW to resolve the problem.

Boot and Log on to the Boot Node

At this point, boot and log on to the boot node on the Cray system. The remaining procedures require dedicated Cray system time.



WARNING: Before starting this procedure, verify that the boot root and shared root file systems are no longer mounted on the SMW. Mounting the file systems on the SMW and boot node simultaneously can corrupt the file systems.

1. Log on to the SMW as `crayadm`.
2. In a shell window, use the `xtbootsys` command to boot the boot node.

```
crayadm@smw:~> xtbootsys
```

3. The `xtbootsys` command displays a series of questions. Cray recommends answering yes by typing `y` in response to each question.

The session pauses at:

```

0) boot bootnode ...
1) boot sdb ...
2) boot compute ...
3) boot service ...
4) boot all (not supported) ...
5) boot all_comp ...
6) boot all_serv ...
10) boot bootnode and wait ...
11) boot sdb and wait ...
12) boot compute and wait ...
13) boot service and wait ...
14) boot all and wait (not supported) ...
15) boot all_comp and wait ...
16) boot all_serv and wait ...
17) boot using a loadfile ...
18) turn console flood control off ...
19) turn console flood control on ...
20) spawn off the network link recovery daemon (xtnlrd)...
q) quit.
Enter your boot choice:

```

Choose option 10 to boot the boot node and wait.

To confirm your selection, press the Enter key or type `y`.

```

Do you want to boot the boot node ? [Yn] Y
Do you want to send the ec_boot event ('no' means to only load memory) ? [Yn] Y
Enter a space-separated list of additional boot parameters (or nothing to do
nothing): <Press the Enter key.>

```

- When the boot node has finished booting, this prompt appears again: `Enter your boot choice`. Do not close the `xtbootsys` terminal session, and do not respond to the prompt at this point in the installation process.

Use this window later to boot the SDB, service, and compute nodes. If the window gets closed, restart `xtbootsys` by using the `-s` option. For more information, see the `xtbootsys(8)` man page.

- Open another shell window on the SMW and use the `ssh` command to log on to the boot node.

```

smw:~ # ssh root@boot
boot:~ #

```

NOTE: The first time that the `root` and `crayadm` accounts on the SMW use the `ssh` command to log on to the boot node, the host key for the boot node is cached. For an initial installation of the boot root on an SMW that has had prior use, it is possible to get the following error message. If this situation is not corrected for the `crayadm` account, an attempt to boot by using `xtbootsys` and a boot automation file may result in a partial failure.

```

@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@  WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!  @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle
attack)!
It is also possible that the RSA host key has just been changed.
The fingerprint for the RSA key sent by the remote host is

```

```
87:65:39:4e:76:de:43:f0:47:f1:d3:12:ac:b7:b0:92.
Please contact your system administrator.
Add correct host key in /root/.ssh/known_hosts to get rid of this message.
Offending key in /root/.ssh/known_hosts:4
RSA host key for boot has changed and you have requested strict checking.
Host key verification failed.
```

If the preceding warning appears when using the `ssh` command to log on to the boot node as `root`, use one of two fixes for the problem.

The first way is to remove the boot node keys using the following commands:

```
smw:~# smw:~# ssh-keygen -R boot
smw:~# su - crayadm
crayadm@smw> ssh-keygen -R boot
crayadm@smw> exit
```

The second way to fix the problem is to edit `/root/.ssh/known_hosts` and `/home/crayadm/.ssh/known_hosts` to remove the previous SSH host key for "boot." The hostname and the IP address are first on the line for the SSH host keys in the `known_hosts` file. The warning lists the line that contains the `ssh` mismatched host key. In the previous example, the `known_hosts` file has an error in line 4.

Change Passwords on the Boot and Service Nodes

For security immediately change the `root` and `crayadm` passwords immediately after login on the boot node for the first time.

1. To change the passwords on the boot node, type the following commands. You are prompted to enter and confirm new root and administrative passwords.

```
boot:~ # passwd root
boot:~ # passwd crayadm
```

2. To change the passwords on the service nodes, type the following commands. Again, you are prompted to enter and confirm new root and administrative passwords.

NOTE: Because the SDB has not been started, use the `-x /etc/opt/cray/sdb/node_classes` option with `xtopview` to specify node/class relationships.

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes
default:// # passwd root
default:// # passwd crayadm
default:// # exit
```

You are prompted to type `c` and enter a brief comment describing the changes you made. To complete your comment, type `Ctrl-d` or a period on a line by itself. Do this each time you exit `xtopview` to log a record of revisions into an RCS system.

Change the Root Password on Compute Nodes

Update the `root` password in the shadow password file on the SMW.

NOTE: To make these changes for a system partition, rather than for the entire system, replace the path specified in the following commands, `/opt/xt-images/templates/default`, with `/opt/xt-images/templates/default-pN`, where *N* is the partition number.

1. Copy the master shadow password file to the template directory.

```
smw:~# cp /opt/xt-images/master/default/etc/shadow \
/opt/xt-images/templates/default/etc/shadow
```

2. Edit the shadow file to include a new encrypted password for root.

```
smw:~# vi /opt/xt-images/templates/default/etc/shadow
```

NOTE: To use the `root` password you created in [Change Passwords on the Boot and Service Nodes](#) on page 106, copy the second field of the root entry in the `/etc/shadow` file on the boot node.

3. Update the boot image to include these changes; follow the steps in [Prepare Compute and Service Node Boot Images](#) on page 100.

NOTE: Several procedures in this chapter include a similar step. You can defer this step and update the boot image once before you [Finish Booting the System](#) on page 132.

Modify SSH Keys for Compute Nodes

The `dropbear` RPM is provided with the CLE release. Using `dropbear` SSH software, an administrator can supply and generate site-specific SSH keys for compute nodes in place of the keys provided by Cray.

Follow these steps to replace the RSA™ and DSA/DSS keys provided by the CLEinstall program.

1. Load the `dropbear` module.

```
crayadm@smw:~> module load dropbear
```

2. Create a directory for the new keys on the SMW.

```
crayadm@smw:~> mkdir dropbear_ssh_keys
crayadm@smw:~> cd dropbear_ssh_keys
```

3. Generate a `dropbear` compatible RSA key.

```
crayadm@smw:~/dropbear_ssh_keys> dropbearkey -t rsa -f ssh_host_rsa_key.db
Will output 1024 bit rsa secret key to 'ssh_host_rsa_key.db'
Generating key, this may take a while...
Public key portion is:
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAQgCQ9ohUgsrrBw5GNk7w2H5RcaBGajmUv8XN6fxg/
YqrsL4t5
CIkNghI3DQDxoiuC/ZVIJctdwZLQJe708eiZee/tg5y2g8JIb3stg+o1/
9BLPDLMeX24FBhCweUpfGCO6Jfm4
Xg4wjKJIGrcmtDJAYoCRj0h9IrdDXXjps7eI4M9XYZ
Fingerprint: md5 00:9f:8e:65:43:6d:7c:c3:f9:16:48:7d:d0:dd:40:b7
```

4. Generate a `dropbear` compatible DSS key.

```
crayadm@smw:~/dropbear_ssh_keys> dropbearkey -t dss -f ssh_host_dss_key.db
Will output 1024 bit dss secret key to 'ssh_host_dss_key.db'
```

Generating key, this may take a while...

Public key portion is:

ssh-dss

```
AAAAB3NzaC1kc3MAAACBAMEkThlE9N8iczLpfg0wUtuPtPcpIs7Y4KbG3Wg1T4CAEXDnfmCKSyuCy
21TMAvVGCvYd80zPtL04ycleUtD5RqEKy0h8jSBs0huEvhaJGHx9FzKfGhWi1ZOVX5vG3R+UCOXG
+71wZp3LU
yOcv/U+GWhalTWpUDaRU81MPRLW7rnAAAAFQCEqnqW61bouSORQ52d
+MRiwp27MwAAAEIAho69yAfGrNzxEI/
kjjyDE5IaxjJpIBF262N9UsxleTX6F650jNoL84fcKq1SL6NV5XJ5000SKgTuVZjpxO913q9SEhkcIOZy
0vRQ8
H5x3osZZ+Bq20QWof+CtWTqCoWN2xvne0NtET4lg81qCt/KGRq1tY6WG
+a01yrvunzQuafQAAACASXvs8h8AA
EK+3TEDj57rBRV4pz5JqWSlUaZStSQ2wJ3Oy1pIJThKfqGWyTv/
nSOWnr8YbQbvH9k1BsyQU8sOc5IjyCFu7+
Exomlyrxq/oirfeSgg6xC2rodcs+jH/K8EKoVtTak3/jHQeZWijRok4xDxwHdZ7e3l2HgYbZLmA5Y=
Fingerprint: md5 cd:a0:0b:41:40:79:f9:4a:dd:f9:9b:71:3f:59:54:8b
```

- As `root`, copy the SSH keys to the boot image template.
To make these changes for a system partition, rather than for the entire system, replace `/opt/xt-images/templates` with `/opt/xt-images/templates-p/N`, where `N` is the partition number.

For the RSA key:

```
crayadm@smw:~/dropbear_ssh_keys> su root
smw:/home/crayadm/dropbear_ssh_keys # cp -p ssh_host_rsa_key.db /opt/xt-images/
templates/default/etc/ssh/ssh_host_rsa_key
```

For the DSA/DSS key:

```
crayadm@smw:~/dropbear_ssh_keys> su root
smw:/home/crayadm/dropbear_ssh_keys # cp -p ssh_host_dss_key.db /opt/xt-images/
templates/default/etc/ssh/ssh_host_dss_key
```

- Update the boot image to include these changes; follow the steps in [Prepare Compute and Service Node Boot Images](#) on page 100.

Modify the `/etc/hosts` File

NOTE: The steps in this section are not required for installation of CLE software.

Site-specific networking requirements can make changes to the `/etc/hosts` file necessary. Follow this procedure to edit the hosts file for the boot node, service nodes, and CNL compute nodes.

IMPORTANT: `CLEinstall` modifies Cray system entries in `/etc/hosts` each time you update or upgrade your CLE software. For additional information, see [Maintain Node Class Settings and Hostname Aliases](#) on page 76.

- Edit the `/etc/hosts` file on the boot node and make site-specific changes.

```
boot:~ # vi /etc/hosts
```

- Copy the edited file to the shared root by using `xtopview` in the default view.

NOTE: Because the SDB has not been started, use the `-x /etc/opt/cray/sdb/node_classes` option with `xtopview` to specify node/class relationships.

```
boot:~ # cp -p /etc/hosts /rr/current/software
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes
default:/ # cp -p /software/hosts /etc/hosts
default:/ # exit
```

3. Make the site-specific changes to the `/opt/xt-images/templates/default/etc/hosts` file on the SMW.

NOTE: To make these changes for a system partition, rather than for the entire system, replace `/opt/xt-images/templates` with `/opt/xt-images/templates-pN`, where *N* is the partition number.

```
smw:~# vi /opt/xt-images/templates/default/etc/hosts
```

- a. Update the boot image to include these changes; follow the steps in [Prepare Compute and Service Node Boot Images](#) on page 100.

NOTE: You can defer this step and update the boot image once before you [Finish Booting the System](#) on page 132.

Configure Login Nodes and Other Network Nodes

Follow these procedures to configure network access and other login class specific information for the login nodes. These procedures also apply to other service nodes, such as network nodes or nodes acting as RSIP servers, which use the shared root and have Ethernet interfaces.



CAUTION: Login nodes and other service nodes do not have swap space. If users consume too many resources, service nodes can run out of memory. When an out of memory condition occurs, the node can become unstable or may crash. System administrators should take steps to manage system resources on service nodes. For example, resource limits can be configured by using the `pam_limits` module and the `/etc/security/limits.conf` file. For more information, see the `limits.conf(5)` man page.

Configure Network Settings for All Login and Network Nodes

The login and network nodes are the portals between the customer's network and the Cray system. Configure basic network information for each login and network node.

1. Use `xtopview` to access each node by either integer node ID or physical ID. For example, to access node 8, type the following:

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes \
-m "network settings" -n 8
```

NOTE: Because the SDB has not been started, use the `-x /etc/opt/cray/sdb/node_classes` option to specify node/class relationships.

TIP: Optionally specify the `-m` option with a brief site-specific comment describing the changes you are making. If this option is specified, `xtopview` does not prompt for comments. This option is suggested when multiple files are changed within a single `xtopview` session.

2. Create and specialize the `/etc/sysconfig/network/ifcfg-eth0` file for the node.

```
node/8:/ # touch /etc/sysconfig/network/ifcfg-eth0
node/8:/ # xtspec -n 8 /etc/sysconfig/network/ifcfg-eth0
```

For a description of specialization, see the `shared_root(5)` man page.

3. Edit `/etc/sysconfig/network/ifcfg-eth0` for the node to include site dependent information. For example, if the site uses static IP addresses, the file might contain the following:

```
BOOTPROTO='static'
STARTMODE='auto'
IPADDR='172.30.12.71/24'
```

Where `"/24"` on the `IPADDR` line is the `PREFIXLEN`, or number of bits that form the network address; alternatively, you may specify `PREFIXLEN` on its own line, although any value appended to `IPADDR` takes precedence. Previous CLE release's `ifcfg` configuration files also may contain the parameter `DEVICE`. Refer to the `ifcfg(5)` man page for more information.

IMPORTANT: If you are configuring an RSIP gateway, you must disable GRO in the `ETHTOOL_OPTIONS` of the network interface by adding the following line:

```
ETHTOOL_OPTIONS="-K iface gro off"
```

Repeat the previous steps in this procedure for each login and network node you have configured.

4. Optional: Specialize and edit `/etc/hosts.allow` and `/etc/hosts.deny` to configure host access control. The information in these files is site dependent. For information about the contents of these files see the `hosts_access(5)` man page. For example, to specialize these files for a single login node, type the following commands.

```
node/8:/ # xtspec -n 8 /etc/hosts.allow
node/8:/ # vi /etc/hosts.allow
node/8:/ # xtspec -n 8 /etc/hosts.deny
node/8:/ # vi /etc/hosts.deny
```

5. Optional: Specialize and edit `/etc/HOSTNAME`. This file is given a value from the `node_class_login_hostname` variable in `CLEinstall.conf`, but may be modified for site-specific considerations.

```
node/8:/ # xtspec -n 8 /etc/HOSTNAME
node/8:/ # vi /etc/HOSTNAME
```

6. Exit from `xtopview`.

```
node/8:/ # exit
```

Configure Class-Specific Login and Network Node Information

After you have configured the basic network information, follow this procedure to configure class-specific information for login or network nodes. The following examples configure the `login` class. Repeat the steps in this procedure for each site-defined class that contains network or RSIP server nodes.

1. Use the `xtopview` command to access login nodes by class. Because the SDB has not been started, use the `-x /etc/opt/cray/sdb/node_classes` option with `xtopview` to specify node/class relationships.

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes \
-m "login class settings" -c login
```

2. Specialize and modify the network configuration file using information from the SMW. Make changes consistent with your network.

```
class/login:/ # xtspec -c login /etc/sysconfig/network/config
class/login:/ # vi /etc/sysconfig/network/config
```

Modify the following variables:

```
NETCONFIG_DNS_STATIC_SEARCHLIST=""
NETCONFIG_DNS_STATIC_SERVERS=""
NETCONFIG_DNS_FORWARDER=""
```

3. Create and specialize the network routes file for the login class. Use information from the SMW to make changes consistent with your network.

```
class/login:/ # touch /etc/sysconfig/network/routes
class/login:/ # xtspec -c login /etc/sysconfig/network/routes
class/login:/ # vi /etc/sysconfig/network/routes
```

4. Create and specialize the `/etc/resolv.conf` file for the login class. Invoke the `netconfig` command to populate the file.

```
class/login:/ # touch /etc/resolv.conf
class/login:/ # xtspec -c login /etc/resolv.conf
class/login:/ # netconfig update -f
```

5. Specialize and edit `/etc/pam.d/sshd` for the login class. To configure PAM to prevent users with key-based authentication from logging in when `/etc/nologin` exists, add the following line from the example below:

IMPORTANT: This must be the first line in the file.

```
class/login:/ # xtspec -c login /etc/pam.d/sshd
class/login:/ # vi /etc/pam.d/sshd
account required pam_nologin.so
```

6. Optional: The following services are turned off by default. Depending on your site policies and requirements, you may need to turn them on by using the `chkconfig` command.

```
cron (see Configure cron Services on page 125)
boot.localnet
flexlm
postfix
```

NOTE: If `postfix` is configured and run on a service node, change the following setting in `/etc/sysconfig/mail` from:

```
MAIL_CREATE_CONFIG="yes" to
MAIL_CREATE_CONFIG="no"
```

Doing so prevents the `master.cf` and `main.cf` postfix configuration files from being recreated during software updates or fixes.

7. Exit `xtopview`.

```
class/login:/ # exit
```

Configure OpenFabrics InfiniBand

InfiniBand is an efficient, low-cost transport between Cray's internal High-speed Network (HSN) and external I/O devices. It can replace or complement Gigabit Ethernet (GigE). OFED/IB driver support is included in the CLE release; OFED and InfiniBand RPMs are installed by default.

You must have the appropriate Host Channel Adapter (HCA) installed for OFED/IB to function correctly. Configure OFED/IB for the particular functionality that you desire. InfiniBand can be configured as follows:

- IB connected service nodes on a Cray system, acting as Lustre servers, to external storage devices. These nodes are commonly referred to as LNET routers. Follow [Configure InfiniBand on Service Nodes](#) on page 112.
- IB can provide IP connectivity between devices on the fabric. To configure IP over InfiniBand (IPoIB), follow [Configure IP Over InfiniBand \(IPoIB\) on Cray Systems](#) on page 113.
- If you are using devices that require the SCSI RDMA Protocol (SRP), follow [Configure and enable SRP on Cray Systems](#) on page 114.

IMPORTANT: Direct-attached InfiniBand file systems require SRP; Lustre file systems external to the Cray system do not require SRP.

For additional information, see *Managing System Software for the Cray Linux Environment (S-2393)*.

Configure InfiniBand on Service Nodes

InfiniBand includes the core OpenFabrics stack and a number of upper layer protocols (ULPs) that use this stack. Configure InfiniBand by modifying `/etc/sysconfig/infiniband` for each IB service node.

1. Use the `xtopview` command to access service nodes with IB HCAs.

For example, if the service nodes with IB HCAs are part of a node class called `lnet`, type the following command:

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes -c lnet
```

Or

Access each IB service node by specifying either a node ID or physical ID. For example, access node 27 by typing the following:

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes -n 27
```

2. Specialize the `/etc/sysconfig/infiniband` file:

```
node/27:/ # xtspec -n 27 /etc/sysconfig/infiniband
```

3. Add IB services to the service nodes by using standard Linux mechanisms, such as executing the `chkconfig` command while in the `xtopview` utility or executing `/etc/init.d/openibd start | stop | restart` (which starts or stops the InfiniBand services immediately). Use the `chkconfig` command to ensure that IB services are started at system boot.

```
node/27:/ # chkconfig --force openibd on
```

4. While in the `xtopview` session, edit `/etc/sysconfig/infiniband` and make these changes.

```
node/27:/ # vi /etc/sysconfig/infiniband
```

- a. By default, IB services do not start at system boot. Change the `ONBOOT` parameter to `yes` to enable IB services at boot.

```
ONBOOT=yes
```

- b. By default at boot time, the Internet Protocol over InfiniBand (IPoIB) driver loads on all nodes where IB services are configured. Verify that the value for `IPOIB_LOAD` is set to `yes` to enable IPoIB services.

```
IPOIB_LOAD=yes
```

IMPORTANT: LNET routers use IPoIB to select the paths that data will travel via RDMA.

- c. The SCSI RDMA Protocol (SRP) driver loads by default on all nodes where IB services are configured to load at boot time. If the Cray system needs SRP services, verify that the value for `SRP_LOAD` is set to `yes` to enable SRP.

```
SRP_LOAD=yes
```

IMPORTANT: Direct-attached InfiniBand file systems require SRP; Lustre file systems external to the Cray system do not require SRP.

5. Exit `xtopview`.

```
node/27:/ # exit
boot:~ #
```

NOTE: The system administrator is prompted to type `c` and enter a brief comment describing the changes made. To complete the comment, type `Ctrl-d` or a period on a line by itself. Do this each time `xtopview` is exited to log a record of revisions into an RCS system.

6. Proper IPoIB operation requires additional configuration. See [Configure IP Over InfiniBand \(IPoIB\) on Cray Systems](#) on page 113.

Configure IP Over InfiniBand (IPoIB) on Cray Systems

1. Use `xtopview` to access each service node with an IB HCA by specifying either a node ID or physical ID. For example, to access node 27, type the following:

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes -n 27
```

2. Specialize the `/etc/sysconfig/network/ifcfg-ib0` file.

```
node/27:/ # xtspec -n 27 /etc/sysconfig/network/ifcfg-ib0
```

3. Modify the site-specific `/etc/sysconfig/network/ifcfg-ib0` file on each service node with an IB HCA.

```
node/27:/ # vi /etc/sysconfig/network/ifcfg-ib0
```

For example, to use static IP address, `172.16.0.1`, change the `BOOTPROTO` line in the file.

```
BOOTPROTO='static'
```

Add the following lines to the file.

```
IPADDR='172.16.0.1'
NETMASK='255.128.0.0'
```

To configure the interface at system boot, change the `STARTMODE` line in the file.

```
STARTMODE='onboot'
```

- Optional: To configure IPoIB for another IB interface connected to this node, repeat step 2 on page 113 and step 3 on page 113 for `/etc/sysconfig/network/ifcfg-ibn`. For LNET traffic, each IB interface should be assigned a unique IP address from the subnet that it will operate on. For TCP/IP traffic, multiple IB interfaces on a node must be assigned unique IP addresses from different subnets.

Configure and enable SRP on Cray Systems

- Use the `xtopview` command to access service nodes with IB HCAs.

For example, if the service nodes with IB HCAs are part of a node class called `ib`, type the following command:

```
boot:~ # xtopview -x /etc/opt/cray/sdb/node_classes -c ib
```

- Edit `/etc/sysconfig/infiniband`

```
ib:// # vi /etc/sysconfig/infiniband and change the value of SRP_DAEMON_ENABLE
to yes:
```

```
SRP_DAEMON_ENABLE=yes
```

- Edit `srp_daemon.conf` to increase the maximum sector size for SRP.

```
ib:// # vi /etc/srp_daemon.conf
```

```
a      max_sect=8192
```

- Optional: Edit `/etc/modprobe.conf.local` to increase the maximum number of gather-scatter entries per SRP I/O transaction.

```
ib:// # vi /etc/modprobe.conf.local
```

```
options ib_srp srp_sg_tablesize=255
```

- Exit from `xtopview`.

```
ib:// # exit
boot:~ #
```

Complete Configuration of the SDB

This procedure proceeds uninterrupted from the previous procedure. At this time, you have three shell sessions open: one running a `tail` command, one running an `xtbootsys` session, and one logged on to the boot node as `root`. The `xtbootsys` session should be paused at the following prompt:

```
0) boot bootnode ...
1) boot sdb ...
2) boot compute ...
3) boot service ...
4) boot all (not supported) ...
5) boot all_comp ...
6) boot all_serv ...
10) boot bootnode and wait ...
11) boot sdb and wait ...
12) boot compute and wait ...
13) boot service and wait ...
14) boot all and wait (not supported) ...
15) boot all_comp and wait ...
16) boot all_serv and wait ...
17) boot using a loadfile ...
18) turn console flood control off ...
19) turn console flood control on ...
20) spawn off the network link recovery daemon (xtnlrd)...
q) quit.
Enter your boot choice:
```

Boot and Configure the SDB Node

Continue in the terminal session for `xtbootsys` that you started in [Boot and Log on to the Boot Node](#) on page 104.

1. Select option 11 to boot the SDB and wait.

To confirm your selection, press the Enter key or type `y`.

```
Do you want to boot the sdb node ? [Yn] Y
Do you want to send the ec_boot event ('no' means to only load memory) ? [Yn] Y
Enter a space-separated list of additional boot parameters (or nothing to do
nothing): <Press the Enter key.>
```

NOTE: Until you start the SDB MySQL database in later in the installation process, a number of error messages similar to `cpadb_mysql_connect: sdb connection failure` may display in your console log file. You may safely ignore these messages.

2. When the SDB node has finished booting, you are prompted to `Enter your boot choice` again. Do not close the `xtbootsys` terminal session. You will use it later to boot the remaining service nodes.
3. In another terminal session, run the `shell_bootnode_first.sh` script on the boot node. This script creates `ssh` keys for `root` on the boot node and copies the `shell_sdbnode_first.sh` script to the SDB.

```
boot:~ # /var/opt/cray/install/shell_bootnode_first.sh
```

This command generates `ssh` DSA, RSA, and ECDSA keys for the `root` account on the boot node.

The `shell_bootnode_first.sh` script copies the `shell_sdbnode_first.sh` script to the SDB node for the next step.

You are prompted to choose the *passphrase* for the `ssh` keys of the root account on the boot node. Use the default file name and specify a null *passphrase*. A null *passphrase* is required to allow passwordless `pdsh` access from the boot node to the other service nodes. This functionality is required by several CLE system utilities, for example `xtshutdown` and Lustre startup on Cray XE and Cray XK systems.

Press the Enter key to choose the defaults and a null *passphrase*.

For the DSA key:

```
Generating public/private dsa key pair.
Enter file in which to save the key (/root/.ssh/id_dsa):
Created directory '/root/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /root/.ssh/id_dsa.
Your public key has been saved in /root/.ssh/id_dsa.pub.
```

For the RSA key:

```
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /root/.ssh/id_rsa.
Your public key has been saved in /root/.ssh/id_rsa.pub.
```

For the ECDSA key:

```
For the ECDSA key:
Generating public/private ecdsa key pair.
Enter file in which to save the key (/root/.ssh/id_ecdsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /root/.ssh/id_ecdsa.
Your public key has been saved in /root/.ssh/id_ecdsa.pub.
```

4. Run the script to install and configure the MySQL database.

Log on to the SDB node. Run the `shell_sdbnode_first.sh` script, and then log off the SDB node.

Press the Enter key to enter a null password when you are prompted for a password.

```
boot:~ # ssh root@sdb
sdb:~ # /tmp/shell_sdbnode_first.sh
Script output
...
Enter password:

Script output
...
sdb:~ # exit
```

5. On the boot node, run the `shell_bootnode_second.sh` script. This script starts the SDB database and completes SDB configuration.

```
boot:~ # /var/opt/cray/install/shell_bootnode_second.sh
```

Change Default MySQL Passwords on the SDB

Access to MySQL databases requires a user name and password. The MySQL accounts and privileges are

MySQL basic	Read access to most tables; most applications use this account
MySQL sys_mgmt	Most privileged; access to all information and commands

For security, Cray recommends changing the default passwords for MySQL database accounts. The valid characters for use in MySQL passwords are:

```
! " # $ % & ' ( )
* + , - . / 0 1 2 3
4 5 6 7 8 9 : ; < =
> ? @ A B C D E F G
H I J K L M N O P Q
R S T U V W X Y Z [
\ ] ^ _ ` a b c d e
f g h i j k l m n o
p q r s t u v w x y
z { | } ~
```

1. If a site-specific MySQL password for root has not been set, complete this step.

- a. Log on to the SDB.

```
boot:~ # ssh root@sdb
```

- b. Invoke the MySQL monitor.

```
sdb:~ # mysql -h localhost -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 4
Server version: 5.5.31-log Source distribution

Type 'help;' or '\h' for help. Type '\c' to clear the buffer.
mysql>
```

- c. Set passwords. Use the actual name of the system database (SDB) node if it is not named `sdb`. For example, the node could be named `sdb-p3` on a partitioned system.

```
mysql> set password for 'root'@'localhost' = password('newpassword');
Query OK, 0 rows affected (0.00 sec)
mysql> set password for 'root'@'%' = password('newpassword');
Query OK, 0 rows affected (0.00 sec)
mysql> set password for 'root'@'sdb' = password('newpassword');
Query OK, 0 rows affected (0.00 sec)
```

2. Optional: Set a site-specific password for other MySQL database accounts.

- a. Change the password for the `sys_mgmt` account. This requires an update to `.my.cnf` in step 4 on page 118.

```
mysql> set password for 'sys_mgmt'@'%' = password('newpassword');
Query OK, 0 rows affected (0.00 sec)
```

- b. Change the password for the `basic` account. This requires an update to `/etc/opt/cray/sysadm/odbc.ini` in step 5 on page 119.

Changing the password for the `basic` MySQL user account will not provide any added security. This read-only account is used by the system to allow all users to run `xtprocdadmin`, `xtnodestat`, and other commands that require SDB access.

```
mysql> set password for 'basic'@'%' = password('newpassword');
Query OK, 0 rows affected (0.00 sec)
```

The connection may time out when making changes to the MySQL database, but it is automatically reconnected. If this happens, messages similar to the following are displayed. These messages may be ignored.

```
ERROR 2006 (HY000): MySQL server has gone away
No connection. Trying to reconnect...
Connection id: 21127
Current database: *** NONE ***

Query OK, 0 rows affected (0.00 sec)
```

3. Exit from MySQL and the SDB.

```
mysql> exit
Bye
sdb# exit
```

4. Optional: If a site-specific password for `sys_mgmt` was set, update the `.my.cnf` file for `root` with the new password. Additionally, update the `.odbc.ini.root` file with the new password.

- a. Edit `.my.cnf` for `root` on the boot node.

```
boot# vi /root/.my.cnf

[client]
user=sys_mgmt
password=newpassword
```

- b. Edit `.my.cnf` for `root` in the shared root.

```
boot# xtopview
default:// # vi /root/.my.cnf

[client]
user=sys_mgmt
password=newpassword
```

- c. Exit `xtopview` and edit `.odbc.ini.root` for the `root` on the boot node. Update **each** database section with the new password.

```
default:// # exit
boot# vi /root/.odbc.ini.root
```

```
Driver          = MySQL_ODBC
Description     = Connector/ODBC Driver DSN
USER            = sys_mgmt
PASSWORD       = newpassword
```

- d. Copy `.odbc.ini.root` to `.odbc.ini`.

```
boot:~ # cp -p /root/.odbc.ini.root /root/.odbc.ini
```

- e. Invoke `xtopview` and edit `.odbc.ini.root` for the root in the shared root. Update **each** database section with the new password.

```
boot# xtopview
default:/: # vi /root/.odbc.ini.root
```

```
Driver          = MySQL_ODBC
Description     = Connector/ODBC Driver DSN
USER           = sys_mgmt
PASSWORD       = newpassword
```

- f. Exit `xtopview` and restart the SDB service on all service nodes.

```
default:/: # exit
boot:~ # pdsh -a /etc/init.d/sdb restart
```

5. Optional: If a site-specific password for `basic` was set, update **each** datasource name in the `/etc/opt/cray/sysadm/odbc.ini` file with the new password. Additionally, update the `/root/.odbc.ini` file with the new password.

- a. Edit `/etc/opt/cray/sysadm/odbc.ini` for the `basic` user on the boot node. Update **each** database section with the new password.

```
boot# vi /etc/opt/cray/sysadm/odbc.ini
```

```
Driver          = MySQL_ODBC
Description     = Connector/ODBC Driver DSN
USER           = basic
PASSWORD       = newpassword
```

- b. Invoke `xtopview` and edit `/root/odbc.ini` in the shared root. Update **each** database section with the new password.

```
boot# xtopview
default:/:# vi /root/odbc.ini
```

```
Driver          = MySQL_ODBC
Description     = Connector/ODBC Driver DSN
USER           = basic
PASSWORD       = newpassword
```

- c. Exit `xtopview`.

Add Node-Specific Services

After the SDB is running, configure the services that run on specific nodes or classes of nodes. The list of supported Cray system services is located in `/etc/opt/cray/sdb/serv_cmd`. An example is provided in `/opt/cray/sdb/default/etc/serv_cmd.example`. Other optional services can be added using this procedure.

1. Invoke the `xtservconfig` command on the boot node to show the available services.

```
boot:~ # xtserveconfig avail
```

Use this command also to show the services already assigned.

```
boot:~ # xtserveconfig list
```

NOTE: Do not add `SYSLOG` or `CRON` by using `xtserveconfig`. Follow [Configure cron Services](#) on page 125 and [Configure System Message Logs](#) on page 131 to configure these services later in the installation process.

2. Assign services to nodes as appropriate. For example, type the following command:

```
boot:~ # xtserveconfig -a add service-name
```

Use the `-c class` option to assign a service to a class of nodes, or the `-n nid` option to assign a service to a specific node.

Configure Additional Services

Boot the login nodes and all other service nodes and configure the following services. Do this before you boot the compute nodes. Note that some of these services are optional.

Configure Service Node MAMU

Service Node MAMU support provides the ability to set aside a small number of repurposed compute nodes for serial workload. These nodes serve as workload management execution nodes, specifically designated for serial workload, and intra-node MPI. The workload manager manages these as standard Linux nodes and support core level placement.

Repurpose Compute Nodes as Service Nodes

CLE and SMW software include functionality to optionally change the role of compute nodes and boot the hardware with service node images. Use this functionality to add service nodes for services that do not require external connectivity, such as `DSL_nodes`. When a compute node is configured with a service node role, that node is referred to as a *repurposed compute node*.

Do not repurpose compute nodes that are intended to be service MAMU nodes until after running the `CLEinstall` program. For more information, see [Configure Service Node MAMU](#) on page 85.

The Cray system hardware state data is maintained in an HSS database where each node is marked with a compute or service node role. By using the `xtcli mark_node` command, you can mark a node in a compute blade to have a role of `service`.

Because they are marked as service nodes within the HSS, repurposed compute nodes are initialized as service nodes by the `CLEinstall` program and are booted automatically when all service nodes are booted.

Configure UFS with Postprocessing Nodes

Use of UFS for user accounts is discouraged. UFS must be used for system accounts like `crayadm`. The directory created provides a minimal administrative account directory for `crayadm`. Configuring the home directory on `/ufs` puts a large load on the Boot RAID and can lead to instability. In this example, `postproc` is a node class created for MAMU nodes and `ufs` is the value of `node_ufs_hostname` in the `CLEinstall.conf` file. Do this task for each defined class that consists of MAMU nodes.

```
boot# xtopview -c postproc
class/postproc:/# xtspec /etc/fstab
```

Add this line:

```
ufs:/ufs/home      /ufs/home      nfs      tcp,rw  0 0
```

```
class/postproc:/# exit
boot# xtopview
default/:/# mkdir -p /ufs/home
default/:/# exit
```

Boot the Remaining Service Nodes

Continue in the terminal session for `xtbootsys` that you started in [Boot and Log on to the Boot Node](#) on page 104.

Boot the login nodes and all other service nodes before booting the compute nodes. After the system is booted, you can reboot it as needed.

1. Select option 13 to boot the service nodes.

```
Enter your boot choice: 13
```

2. Type `p0` to boot the remaining service nodes in the entire system, or `pN` (where *N* is the partition number) to boot a partition. Confirm the selection of nodes and send the `ec_boot` event by pressing the Enter key or typing `Y`.

```
Enter a service list (or nothing to do nothing): p0
'xtcli status -a p0' completed with status 0
'xtcli status -t aries_lcb p0' completed with status 0
Do you want to boot service c0-0c0s0n1,<complete node list not shown>,c0-0c1s2n2 ? [Yn] Y
Do you want to send the ec_boot event ('no' means to only load memory) ? [Yn] Y
Enter a space-separated list of additional boot parameters (or nothing to do nothing): <Press the
Enter key.>
```

3. After the specified service nodes are booted, you are prompted to `Enter your boot choice` again. Do not close the `xtbootsys` terminal session. Use this terminal session later to boot the compute nodes.

Populate the known_hosts File

Run the `shell_ssh.sh` script on the boot node to populate the `known_hosts` file for the root account by using the ssh host keys from the service nodes. Type the following command.

```
boot:~ # /var/opt/cray/install/shell_ssh.sh
```

This is done to verify that `xtshutdown` can contact all service nodes and initiate shutdown procedures by using `pdsh`.

Configure Lustre File Systems

If you plan to configure Lustre file systems, follow the procedures in [Install and Configure Direct-Attached Lustre](#) and then return here to continue the installation.

Create New Login Accounts

NOTE: The steps in this section are not required for installation of CLE software.

To add additional accounts to the shared root for login nodes, use the `groupadd` and `useradd` commands from the default `xtopview` session. For example:

```
boot:~ # xtopview -m "adding user accounts" -c login
class/login:/ # groupadd options
class/login:/ # useradd options
class/login:/ # exit
boot:~ #
```

The `groupadd` and `useradd` commands create group and shadow password entries for new users. However, these commands do not create home directories; create home directories manually. Set the ownership and permissions to enable users to access their home directories. For information about managing user accounts on service nodes, see *Managing System Software for the Cray Linux Environment* (S-2393).

Configure the Login Failure Logging PAM

NOTE: Although the steps in this section are not required for installation of CLE software, Cray recommends that you configure login failure logging on all service nodes.

The `cray_pam` pluggable authentication module (PAM), when configured, provides information to the user at login time about any failed login attempts since their last successful login. To configure this feature, edit the following files on the boot node and then on the service nodes by using the shared root file system:

```
/etc/pam.d/common-auth
/etc/pam.d/common-account
/etc/pam.d/common-session
```

The default location of the `pam_tally` counter file is `/var/log/faillog`. The default location for the `cray_pam` temporary directory is `/var/opt/cray/faillog`. Change these defaults by editing `/etc/opt/cray/pam/faillog.conf` and by using the `file=` option for each `pam_tally` and `cray_pam` entry. You can find an example `faillog.conf` file in `/opt/cray/pam/xtrelease-xtversion/etc`.

Configure `cray_pam` to Log Failed Login Attempts

1. Edit the `/etc/pam.d/common-auth`, `/etc/pam.d/common-account`, and `/etc/pam.d/common-session` files on the boot node.

In these examples, the `pam_faillog.so` and `pam_tally.so` entries can include an optional `file=/path/to/pam_tally/counter/file` argument to specify an alternate location for the tally file.

- a. Edit the `/etc/pam.d/common-auth` file and add the following lines as the first and last entries:

```
boot:~ # vi /etc/pam.d/common-auth

auth required pam_faillog.so [file=alternatepath]
auth required pam_tally.so [file=alternatepath]
```

This example shows the file modified to report failed login using an alternate location for the tally file.

```
#
# /etc/pam.d/common-auth - authentication settings common to all services
#
# This file is included from other service-specific PAM config files,
# and should contain a list of the authentication modules that define
# the central authentication scheme for use on the system
# (e.g., /etc/shadow, LDAP, Kerberos, etc.). The default is to use the
# traditional Unix authentication mechanisms.
```

```
#
auth required pam_faillog.so file=/ufs/logs/tally.log
auth required pam_env.so
auth required pam_unix2.so
auth required pam_tally.so file=/ufs/logs/tally.log
```

- b. Edit the `/etc/pam.d/common-account` file and add the following line as the last entry:

```
boot:~ # vi /etc/pam.d/common-account

account required pam_tally.so [file=alternatepath]
```

This example shows the file modified to report failed login using an alternate location for the tally file.

```
#
# /etc/pam.d/common-account - authorization settings common to all
# services
#
# This file is included from other service-specific PAM config files,
# and should contain a list of the authorization modules that define
# the central access policy for use on the system. The default is to
# only deny service to users whose accounts are expired.
#
account required pam_unix2.so
account required pam_tally.so file=/ufs/logs/tally.log
```

- c. Edit the `/etc/pam.d/common-session` file and add the following line as the last entry:

```
boot:~ # vi /etc/pam.d/common-session

session optional pam_faillog.so [file=alternatepath]
```

This example shows the file modified to report failed login using an alternate location for the tally file.

```
#
# /etc/pam.d/common-session - session-related modules common to all services
#
# This file is included from other service-specific PAM config files,
# and should contain a list of modules that define tasks to be performed
# at the start and end of sessions of *any* kind (both interactive and
# non-interactive). The default is pam_unix2.
#
session required pam_limits.so
session required pam_unix2.so
session optional pam_umask.so
session optional pam_faillog.so file=/ufs/logs/tally.log
```

2. Copy the edited files to the shared root by using `xtopview` in the default view.

```
boot:~ # cp -p /etc/pam.d/common-auth /rr/current/software
boot:~ # cp -p /etc/pam.d/common-account /rr/current/software
boot:~ # cp -p /etc/pam.d/common-session /rr/current/software
boot:~ # xtopview -m "configure login failure logging PAM"
default:/ # cp -p /software/common-auth /etc/pam.d/common-auth
default:/ # cp -p /software/common-account /etc/pam.d/common-account
default:/ # cp -p /software/common-session /etc/pam.d/common-session
```

3. Exit `xtopview`.

```
default/:// # exit
```

Configure the Load Balancer

NOTE: The load balancer service is optional on systems that run CLE.

The load balancer can distribute user logins to multiple login nodes, allowing users to connect by using the same Cray host name, for example `xhostname`.

Two main components are required to implement the load balancer, the `lbname`d service (on the SMW and Cray login nodes) and the site-specific domain name service (DNS).

When an external system tries to resolve `xhostname`, a query is sent to the site-specific DNS. The DNS server recognizes `xhostname` as being part of the Cray domain and shuttles the request to `lbname`d on the SMW. The `lbname`d service returns the IP address of the least-loaded login node to the requesting client. The client connects to the Cray system login node by using that IP address.

The CLE software installation process installs `lbname`d in `/opt/cray-xt-lbname`d on the SMW and in `/opt/cray/lbcd` on all service nodes. Configure `lbname`d by using the `lbname`d.conf and `poller`.conf configuration files on the SMW. For more information about configuring `lbname`d, see the `lbname`d.conf(5) man page.

Configure `lbname`d on the SMW

1. If site-specific versions of `/etc/opt/cray-xt-lbname`d/`lbname`d.conf and `/etc/opt/cray-xt-lbname`d/`poller`.conf do not already exist, copy the provided example files to these locations.

```
smw:~ # cd /etc/opt/cray-xt-lbname/d/
smw:/etc/opt/cray-xt-lbname/d/ # cp -p lbname.d.conf.example lbname.d.conf
smw:/etc/opt/cray-xt-lbname/d/ # cp -p poller.conf.example poller.conf
```

2. Edit the `lbname`d.conf file on the SMW to define the `lbname`d host name, domain name, and polling frequency.

For example, if `lbname`d is running on the host name `smw.mysite.com`, set the login node domain to the same domain specified for the `$hostname`. The Cray system `xhostname` is resolved within the domain specified as `$login_node_domain`.

```
smw:/etc/opt/cray-xt-lbname/d/ # vi lbname.d.conf
```

```
$poller_sleep = 30;
$hostname = "mysite-lb";
$lbname_d_domain = "smw.mysite.com";
$login_node_domain = "mysite.com";
$hostmaster = "rootmail.mysite.com";
```

3. Edit the `poller`.conf file on the SMW to configure the login node names. Because `lbname`d runs on the SMW, `eth0` on the SMW must be connected to the same network from which users log on to the login nodes. Do not put the SMW on the public network.

```
smw:/etc/opt/cray-xt-lbnamed/ # vi poller.conf
```

```
#
# groups
# -----
# login      mycray1-mycray3

mycray1 1 login
mycray2 1 login
mycray3 1 login
```

Install the Load Balancer on an External "White Box" Server

Install `lbnamed` on an external "white box" server as an alternative to installing it on the SMW. **Cray does not test or support this configuration.** A "white box" server is any workstation or server that supports the `lbnamed` service.

1. Shut down and disable `lbnamed`.

```
smw:~# /etc/init.d/lbnamed stop
smw:~# chkconfig lbnamed off
```

2. Locate the `cray-xt-lbnamed` RPM on the Cray CLE 5.0.UP nn Software media and install this RPM on the "white box." Do **not** install the `lbcd` RPM.
3. Follow the instructions in the `lbnamed.conf(5)` man page to configure `lbnamed`, taking care to substitute the name of the external server wherever `SMW` is indicated, then enable the service.

Configure cron Services

NOTE: Configuring `cron` services is optional on CLE systems.

The `cron` daemon is disabled, by default, on the shared root file system and the boot root. It is enabled, by default, on the SMW. Use standard Linux procedures to enable `cron` on the boot root, following [Configure cron for the SMW and the Boot Node](#) on page 125.

On the shared root, configuring `cron` for CLE depends on whether persistent `/var` is set up. If persistent `/var` exists, follow [Configure cron for the Shared Root with Persistent /var](#) on page 126; otherwise, follow [Configure cron for the Shared Root without Persistent /var](#) on page 126.

The `/etc/cron.*` directories include a large number of `cron` scripts. During new system installations and any updates or upgrades, the `CLEinstall` program disables execute permissions on these scripts and they must be manually enabled to be used.

Configure cron for the SMW and the Boot Node

By default, the `cron` daemon on the SMW is enabled and this procedure is required only on the boot node.

1. Log on to the target node as `root` and determine the current configuration status for `cron`.

On the SMW:

```
smw:~# chkconfig cron
cron on
```

On the boot node:

```
boot:~ # chkconfig cron
cron off
```

2. Configure the cron daemon to start.
For this example, enable cron on the boot node:

```
boot:~ # chkconfig --force cron on
```

The `cron` scripts shipped with the Cray customized version of SLES are located under `/etc/cron.hourly`, `/etc/cron.daily`, `/etc/cron.weekly`, and `/etc/cron.monthly`. The system administrator can enable these scripts by using the `chkconfig` command. However, if the system does not have a persistent `/var`, Cray recommends following [Configure cron for the Shared Root without Persistent /var](#).

Configure cron for the Shared Root with Persistent /var

Use this procedure for service nodes by using the shared root on systems that are set up with a persistent `/var` file system.

1. Invoke the `chkconfig` command in the default view to enable the cron daemon.

```
boot:~ # xtopview -m "configuring cron"
default:// # chkconfig --force cron on
```

2. Examine the `/etc/cron.hourly`, `/etc/cron.daily`, `/etc/cron.weekly`, and `/etc/cron.monthly` directories and change the file access permissions to enable or disable distributed cron scripts to meet site needs. To enable a script, invoke `chmod ug+x` to make the script executable. By default, CLEinstall removes the execute permission bit to disable all distributed cron scripts.



CAUTION: Some distributed scripts impact performance negatively on a CLE system. To ensure that all scripts are disabled, type the following:

```
default:// # find /etc/cron.hourly /etc/cron.daily /etc/cron.weekly \
/etc/cron.monthly -type f -follow -exec chmod ugo-x {} \;
```

3. Exit `xtopview`.

```
default:// # exit
```

Configure cron for the Shared Root without Persistent /var

Because CLE has a shared root, the standard cron initialization script `/etc/init.d/cron` activates the cron daemon on all service nodes. Therefore, the cron daemon is disabled by default and must be turned on with the `xtservconfig` command to specify the nodes on which the daemon should run.

1. Edit the `/etc/group` file in the default view to add users who do not have root permission to the "trusted" group. The operating system requires that all cron users who do not have root permission be in the "trusted" group.

```
boot:~ # xtopview
default:/: # vi /etc/group
default:/: # exit
```

2. Create a `/var/spool/cron` directory in the `/ufs` file system on the `ufs` node which is shared among all the nodes of class `login`.

```
boot:~ # ssh root@ufs
ufs:~# mkdir /ufs/cron
ufs:~# cp -a /var/spool/cron /ufs
ufs:~# exit
```

3. Designate a single login node on which to run the scripts in this directory. Configure this node to start `cron` with the `xtservconfig` command rather than the `/etc/init.d/cron` script. This enables users, including `root`, to submit `cron` jobs from any node of class `login`. These jobs are executed only on the specified login node.

- a. Create or edit the following entry in the `/etc/sysconfig/xt` file in the shared root file system in the default view.

```
boot:~ # xtopview
default:/: # vi /etc/sysconfig/xt
```

```
CRON_SPOOL_BASE_DIR=/ufs/cron
```

```
default:/: # exit
```

- b. Start an `xtopview` shell to access all login nodes by class and configure the spool directory to be shared among all nodes of class `login`.

```
boot:~ # xtopview -c login
class/login:/ #
```

- c. Edit the `/etc/init.d/boot.xt-local` file to add the following lines.

```
class/login:/ # vi /etc/init.d/boot.xt-local
```

```
MYCLASS_NID=`rca-helper -i`
MYCLASS=`xtnce $MYCLASS_NID | awk -F: '{ print $2 }' | tr -d [:space:]`
CRONSPPOOL=`xtgetconfig CRON_SPOOL_BASE_DIR`
if [ "$MYCLASS" = "login" -a -n "$CRONSPPOOL" ];then
    mv /var/spool/cron /var/spool/cron.$$
    ln -sf $CRONSPPOOL /var/spool/cron
fi
```

- d. Examine the `/etc/cron.hourly`, `/etc/cron.daily`, `/etc/cron.weekly`, and `/etc/cron.monthly` directories and change the file access permissions to enable or disable distributed `cron` scripts to meet site needs. To enable a script, invoke `chmod ug+x` to make the file executable. By default, `CLEinstall` removes the execute permission bit to disable all distributed `cron` scripts.



CAUTION: Some distributed scripts impact performance negatively on a CLE system. To ensure that all scripts are disabled, type the following:

```
class/login:/ # find /etc/cron.hourly /etc/cron.daily /etc/cron.weekly \
/etc/cron.monthly -type f -follow -exec chmod ugo-x {} \;
```

- e. Exit from the login class view.

```
class/login:/ # exit
```

- f. Enable the `cron` service on a single login node (node 8).

```
boot:~ # xtopview -n 8
node/8:/ # xtserveconfig -n 8 add CRON
node/8:/ # exit
```

The `cron` configuration becomes active on the next reboot. For more information, see the `xtserveconfig(8)` man page.

Configure IP Routes

NOTE: Configuring IP routes for compute nodes is not required on a CLE system.

The `/etc/routes` file can be edited in the CNL template image to provide route entries for compute nodes. This provides a mechanism for administrators to configure routing access from CNL compute nodes to login and network nodes, using external IP destinations without having to traverse RSIP tunnels. Careful consideration should be given before using this capability for general purpose routing.

Configure IP routes

A new `/etc/routes` file is created in the CNL images; it is examined during startup. Non-comment, non-blank lines are passed to the `route add` command. The empty template file contains comments describing the syntax.

NOTE: To make these changes for a system partition, rather than for the entire system, replace `/opt/xt-images/templates/default` with `/opt/xt-images/templates/default-pN`, where *N* is the partition number.

1. Edit `/opt/xt-images/templates/default/etc/routes` and make site-specific changes.
2. Update the boot image to include these changes; follow the steps in [Prepare Compute and Service Node Boot Images](#) on page 100. This step can be deferred by updating the boot image once before you [Finish Booting the System](#) on page 132.

Configure Cray DVS

NOTE: Cray Data Virtualization Service (Cray DVS) is an optional software package.

Cray Data Virtualization Service (Cray DVS) is a parallel I/O forwarding service that enables the transparent use of multiple file systems in CLE systems with close-to-open coherence, much like NFS. DVS provides compute nodes transparent access to external file systems mounted on the service I/O nodes via the Cray high-speed network. Administration of Cray DVS is very similar to configuring and mounting any Linux file system.



CAUTION: DVS service nodes must be dedicated and not share service I/O nodes with other services (e.g., SDB, Lustre, login or MOM nodes).

Cray DVS supports multiple POSIX-compliant, VFS-based file systems. Two supported modes, *serial* and *cluster parallel*, provide functionality for different implementations of existing file systems. Since site conditions and systems requirements differ, please contact your Cray service representative about projecting your preferred file system over DVS.

Because DVS on Cray systems uses the Lustre networking driver (LNET) the following line must be in `/etc/modprobe.conf.local` on DVS servers and in `/etc/modprobe.conf` on DVS clients in those systems:

```
options lnet networks=gni
```

If you configured your system to use a different network identifier than the default (`gni` on Cray systems) you should use that identifier instead. For example, if your LND is configured to use `gni1` as a name, insert the following lines in `modprobe.conf`:

```
options dvsipc_lnet lnd_name=gni1
options lnet networks=gni1
```

Setting the `lnd_name` option for `dvsipc_lnet` is needed so DVS looks for the alternative network identifier since it assumes `gni` as the default. Setting the `networks` option for `lnet` is generally needed when the LNET network type identifier is different.

For Cray service nodes acting only as DVS clients, you will need to insert the following line into the specialized `/etc/modprobe.d/dvs` file for that class of service node:

```
options dvsipc dvsipc_config_type=0
```

For more information, see *Introduction to Cray Data Virtualization Service (S-0005)* and the `dvs(5)` man page.

Configure the System to Mount DVS File Systems

After Cray DVS software has been successfully installed on both the service and compute nodes, you can mount a file system on the compute nodes that require access to the network file system that is mounted on DVS server nodes. When a client mounts the file system, all of the necessary information is specified on the `mount` command.

NOTE: The node that is projecting the file system needs to mount it. Therefore, if the file system is external to the Cray, the DVS server must have external connectivity.

At least one DVS server must be active when DVS is loaded on the client nodes to ensure that all DVS mount points are configured to enable higher-level software, such as the compute node root runtime environment (CNRTE), to function properly.

The following example configures a DVS server at `c0-0c0s4n3` (node 23 on a Cray XE system) to project the file system that is served via NFS from `nfs_serverhostname`. For more information about Cray DVS mount options, see the `dvs(5)` man page.

To make these changes for a system partition, rather than for the entire system, replace `/opt/xt-images/templates/default` with `/opt/xt-images/templates/default-pN`, where *N* is the partition number.

1. Enter `xtopview` with the node view for your DVS server and create the `/dvs-shared` directory that you will be projecting `/nfs_mount` from.

```
boot:~ # xtopview -n 23
node/23:/ # mkdir /dvs-shared
```

2. Specialize the `/etc/fstab` file for the server and add a DVS entry to it.

```
node/23:/ # xtspec -n 23 /etc/fstab
node/23:/ # vi /etc/fstab
```

```
nfs_serverhostname:/nfs_mount /dvs-shared nfs tcp,rw 0 0
node/23:/ # exit
```

3. Log into the DVS server and mount the file system:

```
boot:~ # ssh nid00023
nid00023:/ # mount /dvs-shared
nid00023:/ # exit
```

4. Create mount point directories in the compute image for each DVS mount in the `/etc/fstab` file. For example, type the following command from the SMW:

```
smw:~ # mkdir -p /opt/xt-images/templates/default/dvs
```

5. Optional: Create any symbolic links that are used in the compute node images. For example:

```
smw:~ # cd /opt/xt-images/templates/default
smw:/opt/xt-images/templates/default # ln -s dvs link_name
```

6. To allow the compute nodes to mount their DVS partitions, add an entry in the `/etc/fstab` file in the compute image and add entries to support the DVS mode you are configuring.

```
smw:~# vi /opt/xt-images/templates/default/etc/fstab
```

For serial mode, add a line similar to the following example which mounts `/dvs-shared` from DVS server `c0-0c0s4n3` to `/dvs` on the client node.

```
/dvs-shared /dvs dvs path=/dvs,nodename=c0-0c0s4n3
```

For cluster parallel mode, add a line similar to the following example which mounts `/dvs-shared` from multiple DVS servers to `/dvs` on the client node. Setting `maxnodes` to 1 indicates that each file hashes to only one server from the list.

```
/dvs-shared /dvs dvs path=/dvs,nodename=c0-0c2s1n0:c0-0c2s1n3:c0-0c2s2n0,maxnodes=1
```

For stripe parallel mode, add a line similar to the following example which mounts `/dvs-shared` from the DVS servers to `/dvs` on the client nodes. Specifying a value for `maxnodes` greater than 1 or removing it altogether makes this stripe parallel access mode.

```
/dvs-shared /dvs dvs path=/dvs,nodename=c0-0c2s1n0:c0-0c2s1n3:c0-0c2s2n0
```

For atomic stripe parallel mode, add a line similar to the following example which mounts `/dvs-shared` from the DVS servers to `/dvs` on the client nodes. Specifying `atomic` makes this atomic stripe parallel mode as opposed to stripe parallel mode.

```
/dvs-shared /dvs dvs path=/dvs,nodename=c0-0c2s1n0:c0-0c2s1n3:c0-0c2s2n0,atomic
```

For loadbalance mode, add a line similar to the following example to project `/dvs-shared` from multiple DVS servers to `/dvs` on the client node. The `ro` and `cache` settings specify to mount the data read-only and cache it on the compute node. The `attrcache_timeout` option specifies the amount of time in seconds that file attributes remain valid on a DVS client after they are fetched from a DVS server. Failover is automatically enabled and does not have to be specified.

```
/dvs-shared /dvs dvs path=/dvs,nodename=c0-0c2s1n0:c0-0c2s1n3:c0-0c2s2n0,\
loadbalance,cache,ro,attrcache_timeout=14400
```

7. Update the boot image with the DVS configuration by preparing new compute and service node boot images.

The `CNL_dvs` parameter in the `CLEinstall.conf` file should be enabled before the `CLEinstall` program ran. If DVS was not turned on, edit the `/var/opt/cray/install/shell_bootimage_LABEL.sh` script and set `CNL_DVS=y` before updating the boot image.

NOTE:

It is important to keep `CLEinstall.conf` consistent with changes made to the system configuration in order to avoid unexpected changes during upgrades or updates. Remember to set `CNL_dvs` equal to `yes` in `CLEinstall.conf`.

You can defer updating the boot image and update it once before you [Finish Booting the System](#) on page 132.

Configure System Message Logs

Use the Lightweight Log Management (LLM) System for message log configuration. This system forwards Syslog and other logs to the SMW without keeping a local copy on the Cray system. This system does not normally require additional configuration on the service nodes beyond setting `LLM=yes` in the `CLEinstall.conf` file prior to installing or upgrading the CLE software.

Configure the Node Health Checker

The CLE installation and upgrade processes automatically install and enable the Node Health Checker (NHC) by default; you do not need to change installation parameters or issue any commands. However, you can edit the `/etc/opt/cray/nodehealth/nodehealth.conf` file to specify which NHC tests are to be run and to alter the behavior of NHC tests (including time-out values and actions for tests when they fail); configure time-out values for Suspect Mode and disable/enable Suspect Mode; or disable or enable NHC.

The NHC configuration file, `/etc/opt/cray/nodehealth/nodehealth.conf` is located in the shared root. After you modify the `nodehealth.conf` file, the changes are reflected immediately the next time NHC runs.

To disable NHC entirely, set the value of the `nhcon` global variable in the `nodehealth.conf` file to `off` (the default value is `on`).

Customize Intel Xeon Phi Coprocessor Nodes

The `/opt/xt-images/templates/default` area is used to customize files for compute node `initramfs`. For CLE systems with Intel Xeon Phi Coprocessor nodes, there are some additional customizations that can be done.

1. To customize any file underneath `/opt/xt-images/templates/default` to be different for Intel Xeon Phi coprocessor nodes than for other compute nodes, create a copy of the file with the `.knc` suffix. During the creation of a bootimage with the `shell_bootimage_LABEL.sh` script, any file with the `.knc` suffix will be removed from the `initramfs` for service and compute nodes, but will be renamed for the `initramfs` used to boot the Intel Xeon Phi coprocessor nodes. The `CLEinstall` program configures `/dsl` and writeable `tmp` for the Xeon Phi node (with any differences in syntax) just like they are for the compute node based on settings in the `CLEinstall.conf` file.
2. If the compute nodes mount a Lustre file system, then a special customization is needed for `/opt/xt-images/templates/default/etc/fstab.knc`. Compute nodes mount a file system of type Lustre, but the Intel Xeon Phi coprocessor nodes need to use DVS to access the Lustre file system.

For example, if the compute nodes have this entry in `/opt/xt-images/templates/default/etc/fstab` to mount a Lustre files system from `nid 74` in `/lus/dal`:

```
74@gni:/dal /lus/dal lustre rw,flock,user_xattr 0 0
```

then the `/opt/xt-images/templates/default/etc/fstab.knc` file should change that entry to this one to mount Lustre from a DVS server node (`nid 5`), which is projecting all Lustre file systems underneath the `/lus` mount point:

```
/lus /lus dvs path=/lus,nodename=c0-0c0s1n1,blksize=1048576
```

After changing any files in `/opt/xt-images/templates/default` the boot image will need to be rebuilt. For more information, see [Create Boot Images](#) on page 99.

Finish Booting the System

After all service nodes are booted, boot the compute nodes. After your system is fully booted, you can reboot it as needed. For information about customizing an automatic boot process, see [Configure Boot Automation on the SMW](#) on page 139.

IMPORTANT: If you deferred updating the boot image in any of the previous procedures, update the boot image now by following the steps in [Prepare Compute and Service Node Boot Images](#) on page 100.

Boot the CNL Compute Nodes

At this point in the installation process, all service and login nodes are booted.

1. Return to the terminal session for `xtbootstsys` that you started in [Boot and Log on to the Boot Node](#) on page 104.
2. Select **17** from the `xtbootstsys` menu to boot by using a loadfile. A series of prompts are displayed. Type the responses indicated in the following example. For the `component list` prompt, type `p0` to boot the entire system, or `pN` (where `N` is the partition number) to boot a partition. At the final three prompts, press the `Enter` key.

```
Enter your boot choice: 17
Enter a boot type string (or nothing to do nothing): CNL0
Enter a boot type option (or nothing to do nothing): compute
Enter a component list (or nothing to do nothing): p0
Enter 'any' to wait for any console output,
  or 'linux' to wait for a linux style boot,
  or 'mtk', 'threadstorm', 'ts', or 'xmt' to wait for a MTK style boot,
  or anything else (or nothing) to not wait at all: <Press the Enter key.>
Enter an alternative CPIO archive name (or nothing): <Press the Enter key.>
Do you want to send the ec_boot event ('no' means to only load memory) ? [Yn]
Enter a space-separated list of additional boot parameters (or nothing to do nothing): <Press the
Enter key.>
```

3. The nodes boot, then this prompt is displayed when password-less ssh is not set up. Type `n` to continue.

```
You have not set up password-less access for this account to
execute this command and you also have not provided a password to use.
Answering 'no' to the next question will let you try again.
Do you want to quit this application ? [Yn] n
Enter the default root password:
```

Enter the default root password (`initial0`).

4. After all the compute nodes are booted, return to the xtbootsys menu. Type **q** to exit the xtbootsys program. The output after successful compute node booting is similar to the following:

```

Enter your boot choice: q
'cpio -it -F /tmp/boot/venus-ORANGE-5.2.82.cpio' completed with status 0
'cpio -ivdmu -F /tmp/boot/venus-ORANGE-5.2.82.cpio SNL0/parameters' completed with status 0
Gathering ko files from bootnode:/rr/current/lib/modules/`uname -r`
spawn ssh -o StrictHostKeyChecking=no root@boot-p0 uname -r
Warning: Permanently added 'boot-p0,10.3.1.254' (ECDSA) to the list of known hosts.
3.0.101-0.46.1_1.0502.8871-cray_ari_s
'ssh -o StrictHostKeyChecking=no root@boot-p0 uname -r'
process exited with status '0'
spawn ssh -o StrictHostKeyChecking=no root@boot-p0 file /rr/current/lib/modules/
3.0.101-0.46.1_1.0502.8871-cray_ari_s
/rr/current/lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s:directory
'ssh -o StrictHostKeyChecking=no root@boot-p0 file /rr/current/lib/modules/
3.0.101-0.46.1_1.0502.8871-cray_ari_s'
process exited with status '0'
'rsync -rptgov --copy-unsafe-links --copy-dirlinks --filter "+ */" --filter "+ *.ko"--filter "- *" \
  root@boot-p0:/rr/current/lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s/var/opt/cray/debug/ \
  p0-20150910t094046/shared-root
process exited with status '0'
Gathering ko files from bootnode:/lib/modules/`uname -r`
spawn ssh -o StrictHostKeyChecking=no root@boot-p0 uname -r
3.0.101-0.46.1_1.0502.8871-cray_ari_s
'ssh -o StrictHostKeyChecking=no root@boot-p0 uname -r'
process exited with status '0'
spawn ssh -o StrictHostKeyChecking=no root@boot-p0 file \
  /lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s
/lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s: directory
'ssh -o StrictHostKeyChecking=no root@boot-p0 file \
  /lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s'
process exited with status '0'
'rsync -rptgov --copy-unsafe-links --copy-dirlinks --filter "+ */" --filter "+ *.ko" --filter "- *" \
  root@boot-p0:/lib/modules/3.0.101-0.46.1_1.0502.8871-cray_ari_s/var/opt/cray/debug/ \
  p0-20150910t094046/boot-root'
process exited with status '0'
This session took 3223 seconds (53 minutes, 43 seconds).
#####
Session Boot Summary: 14 nodes completed their boot
#####
INFO: closing exp8
data(config,auto flood control) is off
#####
Your boot session identifier is p0-20150910t094046
#####

```

Flash the nvBIOS for Kepler GPUs

A Cray XC30 system with NVIDIA® Tesla® SXM modules requires an update to the NVIDIA BIOS (nvBIOS) for the NVIDIA K20X and K40s graphics processing units (GPUs). The nvBIOS is unique for each SXM-1 Kepler™ SKU, based on the type of heat sink, as shown below.

GPU Type	Board SKU	Production Firmware Image Version
Kepler K20X (13 fin)	P2085 SKU 202	80.10.44.00.02
Kepler K20X (20 fin)	P2085 SKU 212	80.10.44.00.04
Kepler K20X (30 fin)	P2085 SKU 222	80.10.44.00.05
Kepler K40s (13 fin)	P2085 SKU 209	80.80.4B.00.03
Kepler K40s (20 fin)	P2085 SKU 219	80.80.4B.00.04
Kepler K40s (30 fin)	P2085 SKU 229	80.80.4B.00.05

The CLE software includes a script that automatically determines the SKU version and flashes the nvBIOS with the appropriate firmware.

TIP: You can use the `cselect` command to identify the number and location of the Kepler GPUs. This example shows a system with K20X GPUs on four nodes.

```
login:~# cselect -c "subtype.eq.'nVidia_Kepler'"
4
login:~# cselect -e "subtype.eq.'nVidia_Kepler'"
70-73
```

1. As root on the login node, set the allocation mode for all compute nodes to interactive.

```
login:~# xtprocadmin -km interactive
```

2. Change to the directory containing the NVIDIA scripts.

```
login:~# cd /opt/cray/cray-nvidia/default/bin
```

3. To flash the Kepler K20X GPUs, for example, choose one of the following options.

- To update the entire system:

```
login:~# aprun -n `cselect -c "subtype.eq.'nVidia_Kepler'"` \
-N 1 -L `cselect -e "subtype.eq.'nVidia_Kepler'"` \
./nvFlashBySKU -b
```

- To update a single node or set of nodes:

```
login:~# aprun -n PEs -N 1 -L node_list ./nvFlashBySKU -b
```

NOTE: Replace *PEs* with the number of nodes; replace *node_list* with a comma-separated list of NIDs.

For example, to flash four GPUs on nodes 70-73:

```
login:~# aprun -n 4 -N 1 -L 70,71,72,73 ./nvFlashBySKU -b
c0-0c0s1n0: Nid 70: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n1: Nid 71: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n2: Nid 72: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n3: Nid 73: Successful Cray Graphite K20X nvBIOS flash
```

4. If there is a flash failure, `nvFlashBySKU` displays an error message with the failing node ID, as in this example:

```
c0-0c0s1n3: Nid 73: Failed Cray Graphite K20X nvBIOS flash
```

Depending on the type of failure, `nvFlashBySKU` might display additional information, if available. No flashing is done on unsupported SKUs.

If a GPU fails to flash, the SXM-1 card must be replaced.

5. After flashing is successful, use `xtbootstys` to reboot the nodes from the SMW. For example:

```
crayadm@smw:~> xtbootstys --reboot -L CNL0 -r "rebooting after nvBIOS update" \
c0-0c0s1n0,c0-0c0s1n1,c0-0c0s1n2,c0-0c0s1n3
```

TIP: You can use `xtprocadmin` on the login node to determine each node name from the `cselect` output, as in this example:

```
login:~# xtprocadmin -n `cselect -e "subtype.eq.'nVidia_Kepler'"`
  NID   (HEX)   NODENAME   TYPE     STATUS   MODE
  70    0xf8    c0-0c0s1n0 compute   up       batch
  71    0xf9    c0-0c0s1n1 compute   up       batch
  72    0xfa    c0-0c0s1n2 compute   up       batch
  73    0xfb    c0-0c0s1n3 compute   up       batch
```

- After the reboot is successful, log on to the login node as root and change to the directory containing the NVIDIA scripts.

```
login:~# cd /opt/cray/cray-nvidia/default/bin
```

To verify that all GPUs are reporting the correct nvBIOS version (see the table above), choose one of the following options:

- To display the nvBIOS versions for the entire system:

```
login:~# aprun -n `cselect -c "subtype.eq.'nVidia_Kepler'"` \
-N 1 -L `cselect -e "subtype.eq.'nVidia_Kepler'"`\
./xkcheck -n -c -f | grep Version
```

- To display the nvBIOS versions for a single node or set of nodes:

```
login:~# aprun -n PEs -N 1 -L node_list \
./xkcheck -n -c -f | grep Version
```

Replace *PEs* with the number of nodes; replace *node_list* with a comma-separated list of NIDs.

For example:

```
login:~# aprun -n 4 -N 1 -L 70,71,72,73 \
./xkcheck -n -c -f | grep Version
4 nodes report VBIOS Version           : 80.10.3D.00.05
```

- Reset the compute nodes to the normal batch or interactive mode using the xtprocadmin command.

Test the System for Basic Functionality

- If the system was shut down by using xtshutdown, remove the /etc/nologin file from all service nodes to permit a non-root account to log on.

```
smw:~# ssh root@boot
boot:~ # xtunspec -r /rr/current -d /etc/nologin
```

- Log on to the login node as `crayadm`.

```
boot:~ # ssh crayadm@login
```

- Use system-status commands, such as `xtnodestat`, `xtprocadmin`, and `apstat`.

The `xtnodestat` command displays the current allocation and status of the compute nodes, organized by physical cabinet. The last line of the output shows the number of available compute nodes. The output for Cray XE and Cray XK systems follows.

```
crayadm@login:~> xtnodestat
Current Allocation Status at Fri May 21 07:11:48 2012
```



```

 2      0x2  c0-0c0s1n0  service      up interactive
 4      0x4  c0-0c0s2n0  service      up interactive
 6      0x6  c0-0c0s3n0  service      up interactive
. . .
93     0x5d  c0-0c2s1n3  service      up interactive
94     0x5e  c0-0c2s0n2  service      up interactive
95     0x5f  c0-0c2s0n3  service      up interactive

```

The output for Cray XC30 systems follows.

```

crayadm@login:~> xtprocadmin
  NID  (HEX)  NODENAME  TYPE  STATUS  MODE
  1    0x1    c0-0c0s0n1  service  up    batch
  2    0x2    c0-0c0s0n2  service  up    batch
  5    0x5    c0-0c0s1n1  service  up    batch
  6    0x6    c0-0c0s1n2  service  up    batch
  8    0x8    c0-0c0s2n0  compute  up    batch
  9    0x9    c0-0c0s2n1  compute  up    batch
 10   0xa    c0-0c0s2n2  compute  up    batch

```

The `apstat` command displays the current status of all applications running on the system.

```

crayadm@login:~> apstat -v
Compute node summary
  arch  config  up  resv  use  avail  down
  XT    733    733  107   89   626    0

Total pending applications: 4
Pending Pid  User  w:d:N  NID  Age  Command  Why
 17278  crayadm  1848:1:24  5  0h53m  ./app1  Busy
 17340  crayadm  1848:1:24  5  0h53m  ./app1  Busy
 17469  crayadm  1848:1:24  5  0h52m  ./app1  Busy
 26155  crayadm  1848:1:24  5  0h12m  ./app2  Busy

Total placed applications: 2
  Apid  ResId  User  PEs  Nodes  Age  State  Command
1631095  135  alan-1  64   4  0h31m  run  mcp
1631145  140  flynn  128  8  0h05m  run  TRON-JA307020

```

4. Run a simple job on the compute nodes.

At the conclusion of the installation process, the `CLEinstall` program provides suggestions for runtime commands and indicates how many compute nodes are available for use with the `aprun -n` option.

For `aprun` to work cleanly, the current working directory on the login node should also exist on the compute node. Change your current working directory to either `/tmp` or to a directory on a mounted Lustre file system.

For example, type the following.

```

crayadm@login:~> cd /tmp
crayadm@login:~> aprun -b -n 16 -N 1 /bin/cat /proc/sys/kernel/hostname

```

This command returns the hostname of each of the 16 compute nodes used to execute the program.

```

nid00010
nid00011
nid00012
nid00020
nid00016
nid00040
nid00052

```

```
nid00078
nid00084
nid00043
nid00046
nid00049
. . .
```

5. Test file system functionality. For example, if you have a Lustre file system named `/mylustmnt/filesystem`, type the following.

```
crayadm@login:~> cd /mylustmnt/filesystem
crayadm@login:/mylustremnt/filesystem> echo lustretest > testfile
crayadm@login:/mylustremnt/filesystem> aprun -b -n 5 -N 1 /bin/cat ./testfile
lustretest
lustretest
lustretest
lustretest
lustretest
Application 109 resources: utime ~0s, stime ~0s
```

6. Test the optional features that you have configured on your system.
- a. To test RSIP functionality, log on to an RSIP client node (compute node) and ping the IP address of the SMW or other host external to the Cray system. For example, if `c0-0c0s7n2` is an RSIP client, type the following commands.

```
crayadm@login:~> exit
boot:~ # ssh root@c0-0c0s7n2
root@c0-0c0s7n2's password:
Welcome to the initramfs
# ping 172.30.14.55
172.30.14.55 is alive!
# exit
Connection to c0-0c0s7n2 closed.
boot:~ # exit
```

NOTE: RSIP clients on the compute nodes make connections to the RSIP server(s) during system boot. Initiation of these connections is staggered over a two minute window; during that time, connectivity over RSIP tunnels is unreliable. Avoid using RSIP services for three to four minutes following a system boot.

- b. To check the status of DVS, type the following command on the DVS server node.

```
crayadm@login:~> ssh root@nid00019 /etc/init.d/dvs status
DVS service: ..running
```

To test DVS functionality, invoke the `mount` command on any compute node.

```
crayadm@login:~> ssh root@c0-0c0s7n2 mount | grep dvs
/dvs-shared on /dvs type dvs
(rw,blksize=16384,nodename=c0-0c0s4n3,nocache,nodatasync,\
retry,userenv,clusterfs,maxnodes=1,nnodes=1)
```

Create a test file on the DVS mounted file system. For example, type the following.

```
crayadm@login:~> cd /dvs
crayadm@login:/dvs> echo dvstest > testfile
crayadm@login:/dvs> aprun -b -n 5 -N 1 /bin/cat ./testfile
dvstest
```

```

dvstest
dvstest
dvstest
dvstest
Application 121 resources: utime ~0s, stime ~0s

```

- Following a successful installation, the file `/etc/opt/cray/release/clerelease` is populated with the installed release level. For example,

```

crayadm@login:~> cat /etc/opt/cray/release/clerelease
5.2.UP04

```

If the preceding simple tests ran successfully, the system is operational. Cray recommends using the `xthotbackup` utility to create a backup of a newly updated or upgraded system. For more information, see the `xthotbackup(8)` man page.

Configure Boot Automation on the SMW

A sample boot automation file, `/opt/cray/hss/default/etc/auto.generic.cnl`, is provided as a basis for further customizing the boot process. A system shutdown is required to test the customized boot automation files.

For more information about boot automation, see the `xtbootstys(8)` man page.

- Use your site-specific procedures to shut down the system. For example, to shutdown using an automation file, type the following:

```

crayadm@smw:~> xtbootstys -s last -a auto.xtshutdown
'xtcli shutdown $data(idlist)' completed with status 0

connecting to boot node (boot-p0)

spawn ssh -o StrictHostKeyChecking=no -x root@boot-p0
Warning: Permanently added 'boot-p0' (ECDSA) to the list of known hosts.
Password:

You have not set up password-less ssh from this
account on the SMW to 'root' on the boot node
and you also have not provided a password to use.
** Answering 'no' to the next question will let you try again. **
Do you want to quit this application ? [Yn] n
Enter your root password :

```

Enter the default root password (`initial0`).

Although not the preferred method, alternatively execute these commands as `root` from the boot node to shutdown your system.

```

boot:~ # xtshutdown -y
boot:~ # shutdown -h now;exit

```

- Prepare a boot automation file. If no boot automation file exists, copy the template file.

```

crayadm@smw:~> cp -p /opt/cray/hss/default/etc/auto.generic.cnl \
/opt/cray/hss/default/etc/auto.xthostname

```

- Edit the boot automation file.

```
crayadm@smw:~> vi /opt/cray/hss/default/etc/auto.xthostname
```

NOTE: The boot automation file contains many of the following commands but the lines are commented out. Uncomment the pertinent lines and edit them as needed.

- a. To enable non-root logins following a system shutdown, add the following as the last command:

```
lappend actions { crms_exec_on_bootnode "root" \
"xtunspec -r /rr/current -d /etc/nologin" }
```

- b. If you have configured Lustre file systems for your system, see [Configure a Boot Automation File for DAL](#).
- c. Make additional site-specific changes as needed and save the file.

4. Use the `xtbootsys` command to boot the Cray system.



CAUTION: Shut down the Cray system before invoking the `xtbootsys` command. If installing to an alternate system set, shut down the currently running system before booting the new boot image.

Type the following command to boot the entire system.

```
crayadm@smw:~> xtbootsys -a auto.xthostname
```

Or

Type the following command to boot a partition.

```
crayadm@smw:~> xtbootsys --partition pN -a auto.xthostname
```

5. Reboot the system and confirm that shutdown and boot procedures operate as expected.

The software installation of your Cray system is complete. Cray recommends that you use the `xthotbackup` utility to create a backup of your newly installed system. For more information, see the `xthotbackup(8)` man page.

Configure Boot Automation for SDB Node Failover

If SDB node failover is configured on the system and there are commands in the boot automation script that apply to the SDB, follow these steps to ensure the appropriate boot automation commands are invoked if an SDB node fails.

SDB-specific commands in the boot automation script must be invoked for the backup SDB node in the event of a failover. However, the boot automation script does not apply to the backup SDB node in a failover situation.

1. Create or edit the `sdbfailover.conf` file in the shared root file system in the default view.

```
boot001:~# xtopview
default:/: # vi /etc/opt/cray/sdb/sdbfailover.conf
```

Make optional site-specific changes. For example, if the boot automation file started a batch scheduler (it was not started by using `chkconfig`) or set up a route to an external license server, add the same commands to the `sdbfailover.conf` file so that they are invoked when the backup SDB node is started. For example:

```
#
# Commands to be run on the backup sdb node after it has failed over
#
```

```
/bin/netstat -r
/sbin/route add default gw login
/etc/init.d/torque_server start
/etc/init.d/moab start
```

2. Exit xtopview.

```
default/:// # exit
```

Post Installation System Management

The Cray system running CLE software should now have an operational . For information about additional software you may need on your system, including programming environment and batch software, see [Install Additional Software](#). [Install RPMs](#) provides generic instructions for installing RPM Package Manager (RPM) packages.

For information about additional system administrative tasks to manage operation of your system, see *Managing System Software for the Cray Linux Environment (S-2393)*. It presents the following topics in greater detail:

- Managing the system
- Monitoring system activity
- Managing user access
- Modifying an installed system
- Managing services
- SMW and CLE System Administration Commands

Managing System Software for the Cray Linux Environment (S-2393) provides complete documentation for most CLE features. These features or subsystems may require site-specific configuration and administration.

- Application Level Placement Scheduler (ALPS)
- OpenFabrics Interconnect Drivers
- Node Health Checker (NHC)
- System Environmental Data Collector (SEDC)

If you wish to use the following optional features, additional configuration is required. See *Managing System Software for the Cray Linux Environment (S-2393)* for more information.

- Dynamic Shared Objects and Cluster Compatibility Mode (CCM)
- Comprehensive System Accounting (CSA)

Install SMW Software on the Second SMW

The SMW that is installed second (the *secondary SMW*) will initially become the passive SMW when the SMW HA cluster is fully configured.

Note these requirements for the second SMW:

- The second SMW must run exactly the same version of operating system and Cray SMW software as the first SMW.

- As on the first SMW, two Ethernet ports are used for heartbeat connections between the two SMWs: `eth2` (on the first quad Ethernet card) and `eth4` (on the second quad Ethernet card); see [Network Connections for an SMW HA System](#) on page 16. Note that these ports are marked as "Reserved for SMW failover" in [Network Connections](#) on page 22.
- If you set up the SUSE firewall and IP tables on the first SMW, use the same configuration on the second SMW.

For the second SMW, you must skip several steps when installing the base operating system and SMW software. The following procedures describe how to install this software.

Prepare to Install SMW Software on the Second SMW

Before installing the SMW software, use this procedure to prepare the second SMW.

The SMW that is installed second will initially become the passive SMW when the SMW HA cluster is fully configured. The examples in the installation procedures show the host name `smw2` for the second SMW.

Ensure that the boot RAID is disconnected.

When installing the operating system, only the boot disk should be connected to the SMW. All other internal disks should be disconnected (ejected). The boot RAID **must** be disconnected to prevent data corruption when installing the operating system.

Install the SMW Release Package

Follow these procedures to perform an initial or clean installation of the Cray System Management Workstation (SMW) 7.2.UP04 release package.

Cray provides two rack-mount SMW models: the Dell PowerEdge™ R815 Rack Server and the Dell PowerEdge™ R630 Rack Server. The figure below shows an easy way to distinguish between the two rack-mount models when viewing them from the front.

Figure 5. Distinguishing Features of Dell R815 and R630 Servers



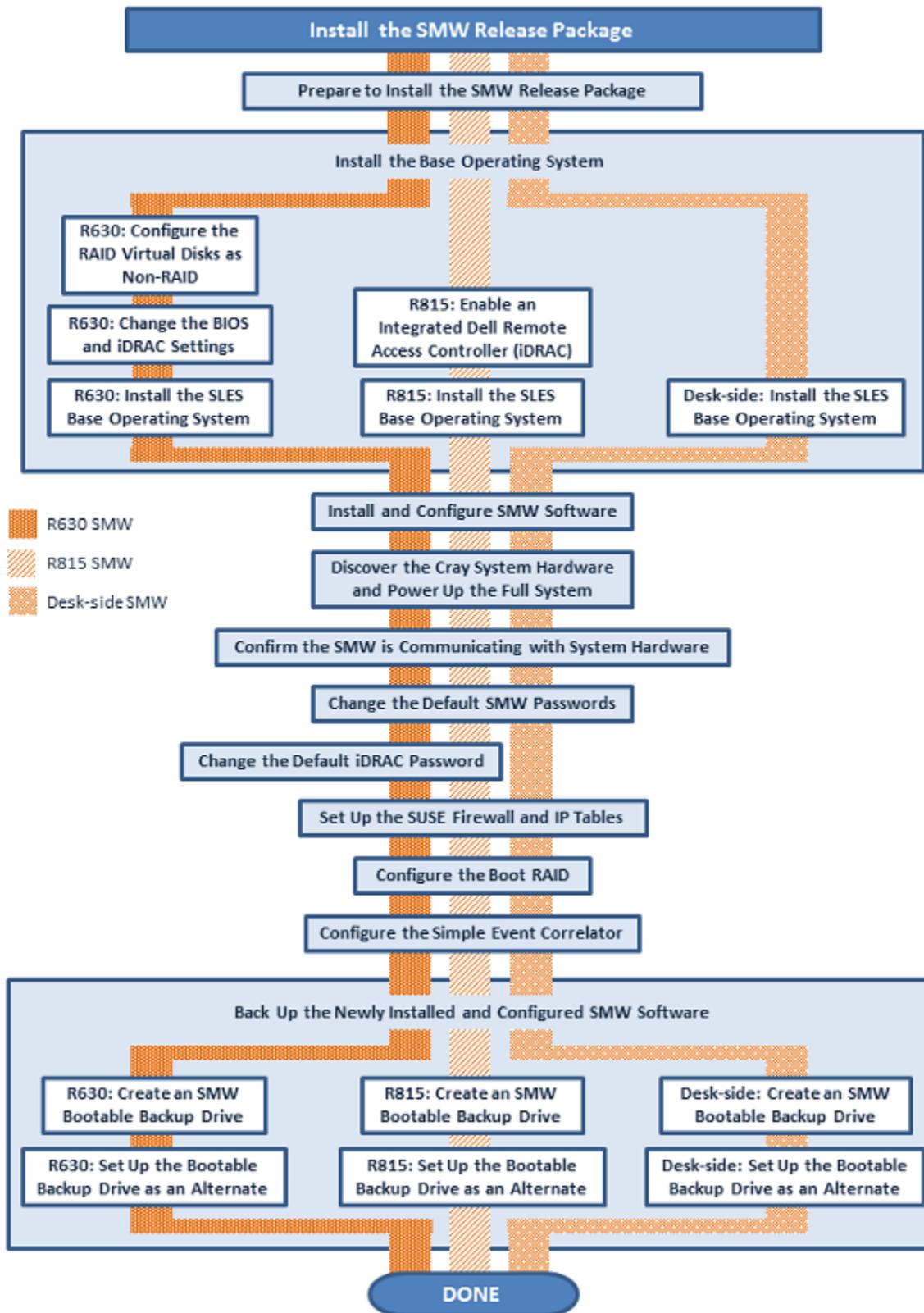
Dell R815: 2U high and 6 drive bays



Dell R630: 1U high and 8 drive bays

Most of the installation procedures that follow apply to all SMW models. Those that apply to only one or more of the models are clearly marked, and the visual guide below also indicates applicability.

Figure 6. Visual Guide to SMW Release Package Installation



Prepare to Install the SMW Release Package

- Read the *SMW Release Errata* and the *SMW README* provided with the SMW release package for any additional installation-related requirements, corrections to this installation guide, and other relevant information about the release package.
- Read the Field Notices (FN) related to kernel security fixes to identify any changes to this release package. Apply any needed changes before installing the new software.
- Use the correct publication. To install the SMW release package on a system configured for SMW high availability (HA) with the SMW failover feature, use *SMW HA XC Installation Guide* (S-0044). Otherwise, use *System Management Workstation (SMW) Software Installation Guide* (S-2480).
- Verify that the network connections are in place (see [Network Connections](#) on page 22).
- Know which configuration values are site-specific and which are defaults (see [Configuration Values](#) on page 22).
- Refer to the default passwords used during the installation process (see [Passwords](#) on page 24).
- Become familiar with the mapping of Linux device names to physical drive slots (see [Rack-mount SMW: Linux /dev Names Mapped to Physical Drive Slots](#)).

Network Connections

The following network connections are required.

- A standalone SMW with a single quad-ethernet card has these private network connections:
 - eth0 - To the customer network
 - eth1 - To the Hardware Supervisory System (HSS) network
 - eth2 - Reserved for SMW failover
 - eth3 - To the boot node
- An SMW configured for SMW failover has a second quad-ethernet card with these connections:
 - eth4 - Used for SMW failover
 - eth5 - Used for mirrored storage for the power management database (PMDB), if available
 - eth6-7 - Reserved for future use

NOTE: Ethernet port assignments are valid only after the SMW software installation completes (see [R815 SMW: Install the SLES Base Operating System](#) on page 24).

- The SMW must have a Fibre Channel or Serial Attached SCSI (SAS) connection to the boot RAID.
- The boot node must have a Fibre Channel or SAS connection to the boot RAID.
- The service database (SDB) node must have a Fibre Channel or SAS connection to the boot RAID.

IMPORTANT: The SMW must be disconnected from the boot RAID before the initial installation of the SLES software. Ensure that the Fibre Channel optic cable connectors or SAS cable connectors have protective covers.

Configuration Values

The following IP addresses are set by default and are not site dependent.

NOTE: These default IP addresses are only for a standalone SMW. For an SMW HA system, see the default IP addresses in [Configuration Values for an SMW HA System](#) on page 16.

Table 16. Default IP Addresses

IP Address	Description
10.1.0.1	Primary boot RAID controller
10.1.0.2	Secondary boot RAID controller
10.1.0.15	Storage RAID controller
10.1.1.1	SMW, eth1
10.2.1.1	SMW, eth2 - Reserved for SMW failover
10.3.1.1	SMW, eth3
10.3.1.254	boot node
10.4.1.1	SMW, eth4 - Reserved for SMW failover
127.0.0.1	localhost (loopback)

The following configuration values are site dependent. Record the actual values for the installation site in the third column. References to rack-mount SMW include both the Dell R815 and Dell R630 models..

NOTE: In addition to these values, there are HA-specific values that apply to an SMW HA system. See [Configuration Values for an SMW HA System](#) on page 16.

Table 17. Site-dependent Configuration Values

Description	Example	Actual Value
SMW hostname	xtsmw	
Domain	cray.com	
Aliases	cray-smw smw01	
Customer network IP address	192.168.78.68	
Customer network netmask	255.255.255.0	
Default gateway	192.168.78.1	
Domain names to search	us.cray.com mw.cray.com	
Nameserver IP address	10.0.73.30 10.0.17.16	
For rack-mount SMW only: iDRAC hostname	cray-drac	
For rack-mount SMW only: iDRAC IP address	192.168.78.69	
For rack-mount SMW only: iDRAC Subnet Mask	255.255.255.0	
For rack-mount SMW only: iDRAC Default GW	192.168.78.1	
Timezone	US/Central	
NTP servers	ntpghost1 ntpghost2	

Description	Example	Actual Value
X dimension	1-64	
Y dimension	Cray XE and Cray XK systems: 1-16; Cray XC Series Systems: 1-32	
Topology Class	Cray XE and Cray XK systems: 0, 1, 2, 3; Cray XC Series Systems: 0, 2 NOTE: Regardless of the number of cabinets in the system, Cray XC Series air-cooled systems must be set to 0. Cray XC Series liquid-cooled systems can be class 0 or 2.	

Passwords

The following default account names and passwords are used throughout the SMW software installation process. Cray recommends changing these default passwords after completing the installation.

Table 18. Default System Passwords

Account Name	Password
root	initial0
crayadm	crayadm
cray-vnc	cray-vnc
mysql	None; a password must be created
admin (DDN™ boot RAID)	password
user (DDN boot RAID)	password
admin (DDN storage RAID)	password
user (DDN storage RAID)	password
root (iDRAC)	initial0

R815 SMW: Install the SLES Base Operating System

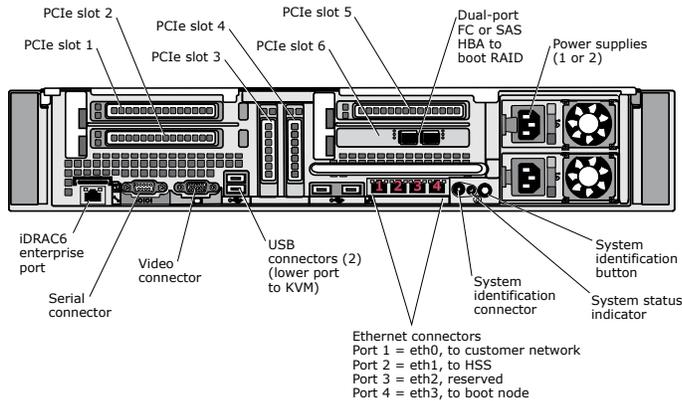
This procedure describes how to install SLES11SP3 as the base operating system on an R815 SMW. Use the DVD labeled `Cray-SMWbase11SP3-` to perform this installation.

1. If the SMW is up, `su` to `root` and shut it down.

```
crayadm@smw> su - root
smw# shutdown -h now;exit
```

2. Disconnect the SMW connection to the boot RAID; disconnect the data cables and place protective covers on the Fibre Channel or SAS cables and connectors (if present).

Figure 7. Dell R815 SMW Rear Connections



3. Eject all the disk drives except for the primary disk in slot 0.
4. Power up the SMW. When the BIOS power-on self-test (POST) process begins, quickly press the F2 key after the following messages appear in the upper-right of the screen.

```
F2 = System Setup
F10 = System Services
F11 = BIOS Boot Manager
F12 = PXE Boot
```

When the **F2** keypress is recognized, the **F2 = System Setup** line changes to **Entering System Setup**.

After the POST process completes and all disk and network controllers have been initialized, the BIOS set-up screen appears.

5. Use the down-arrow key to highlight **Boot Settings**. Press the **Enter** key.

A window listing the following appears:

```
Boot Mode ..... BIOS
Boot Sequence ..... <ENTER>
USB Flash Drive Emulation Type..... <ENTER>
Boot Sequence Retry ..... <Disabled>
```

6. Use the down-arrow key to highlight **Boot Sequence**. Press the **Enter** key.

A window listing the following appears.

```
√ 1. Hard Drive C: (Integrated SAS 500 ID0A LUN0 ATA)
√ 2. Virtual CD
√ 3. Sata Optical Drive
√ 4. Embedded NIC 1 MBA v6.0.11 Slot 0100
√ 5. Virtual Floppy
```

7. Using the up-arrow and down-arrow keys to select, and using the space key to enable/disable entries, modify the list so that only the 1. `Hard Drive C: (Integrated SAS 500 ID0A LUN0 ATA)` entry has a check mark.

```

√ 1. Hard Drive C: (Integrated SAS 500 ID0A LUN0 ATA)
  2. Virtual CD
  3. Sata Optical Drive
  4. Embedded NIC 1 MBA v6.0.11 Slot 0100
  5. Virtual Floppy

```

8. Press the **Esc** key to exit the **Boot Sequence** window.
9. Press the **Esc** key again to exit the **Boot Settings** window.
10. Insert the base operating system DVD labeled `Cray-SMWbase11SP3-` into the CD/DVD drive. (The DVD drive on the front of the SMW may be hidden by a removable decorative bezel.)
11. Press the **Esc** key a final time to save changes and exit the BIOS set up. A screen listing exit options appears.

```

Save changes and exit
Discard changes and exit
Return to Setup

```

12. Ensure that `Save changes and exit` is highlighted. Then press the **Enter** key. The SMW resets automatically.
13. When the BIOS POST process begins again, within 5 seconds, press the **F11** key after the following messages appear in the upper-right of the screen.

```

          F2 = System Setup
          F10 = System Services
          F11 = BIOS Boot Manager
          F12 = PXE Boot

```

When the **F11** keypress is recognized, the `F11 = BIOS Boot Manager` line changes to `Entering BIOS Boot Manager`.

14. Watch the screen carefully as text scrolls until `iDRAC6 Configuration Utility 1.57` appears. Within 5 seconds, press the **Ctrl-E** key when the prompt `Press <Ctrl-E> for Remote Access Setup within 5 sec...` displays.

```

0 5 0 ATA WDC WD5000BPVT-0 1A01 465 GB
LSI Corporation MPT2 boot ROM successfully installed!

```

```

iDRAC6 Configuration Utility 1.57
Copyright 2010 Dell Inc. All Rights Reserved

```

```

iDRAC6 Firmware Revision version: 1.54.15
Primary Backplane Firmware Revision 1.07

```

```

-----
IPv6 Settings
-----

```

```

IPv6 Stack : Disabled

```

```

Address 1      : ::
Default Gateway : ::
-----
      IPv4 Settings
-----
IPv4 Stack      : Enabled
IP Address      : 172. 31. 73.142
Subnet mask     : 255.255.255. 0
Default Gateway : 172. 31. 73. 1
Press <Ctrl-E> for Remote Access Setup within 5 sec...

```

The **iDRAC6 Configuration Utility** window appears.

15. In the **iDRAC6 Configuration Utility**, select **Virtual Media Configuration**.

The **Virtual Media Configuration** window appears.

16. Select the **Virtual Media** line until it indicates **Detached**.

```

Virtual Media ..... Detached

```

17. Press the **Esc** key to exit the **Virtual Media Configuration** window.

18. Press the **Esc** key again to exit the **iDRAC6 Configuration Utility** window.

The **BIOS Boot Manager** screen appears.

19. Use the up-arrow and down-arrow keys to highlight the **SATA Optical Drive** entry.

20. Press the **Enter** key to boot from the installation DVD.

21. Within 10 to 15 seconds after the SUSE Linux Enterprise Server boot menu displays, use the down-arrow key to scroll down and select the **Cray SMW Initial Install Rackmount** option.

```

- Boot from Hard Disk
- Cray SMW Initial Install
- Cray SMW Initial Install Rackmount
- Cray SMW Upgrade Install
- Repair Installed System
- Rescue System
- Check Installation Media

```

Then press the **Enter** key.

If the timeout ends before a selection is made, the system will boot from the hard disk (the default selection). If that happens, shut down the SMW, then begin the power-up sequence again.

As the base installation progresses, the following phrases appear on the screen:

```

Analyzing Computer
System Probing
Preparing System for Automated Installation
Installation Settings

```

After these screens are displayed, the installation pauses on **Installation Settings**.

22. Partition the drive for installation of the base operating system and SMW software.

- a. Remove and recreate the `sda` partitions.

1. For each partition of `sda`, double-click on the Device name of that specific **Hard Disk** table entry, which opens the `Partition: /dev/sd...` table for that drive and which has **Edit**, **Resize** and **Delete** buttons below the table.
 2. Select the **Delete** button. A pop-up window indicating `Really delete /dev/sd...` appears.
 3. Verify the device, and select **Yes** to delete the existing disk partition on that drive. The specific `/dev/sd...` entry will be removed from the displayed list.
 4. Repeat these steps for all entries associated with the `sda` disk until the displayed list is empty.
- b. Create new partitions for `swap` and `root` for the `sda` device sized for this system. This is the target boot device for the operating system installation.

1. Select the **Add** button at the bottom of the **Hard Disk: /dev/sd...** window, which opens the **Add Partition on /dev/sd...** screen.
2. Select **Primary Partition**, then select **Next**. The **New Partition Size** screen displays.
3. Select **Custom** size, then enter the swap partition size into the box. This swap partition size should be the same as the total system RAM memory on the SMW.

Enter either **8GB** or **32GB**, depending on the total system RAM memory on the SMW.

TIP: For an SMW that has one power supply, enter **8GB**. For an SMW that has two power supplies, enter **32GB**.

Then select **Next**. The **Formatting Options and Mounting Options** screen displays.

4. Select **Format partition** and select **Swap** from the pull-down menu under **File system**.
5. Select **Mount partition** and select **Swap** from the pull-down menu under **Mount Point**.
6. Select the **Finish** button. The **Hard Disk: /dev/sd...** window displays again, but with the new partition entry.
7. Select the **Add** button at the bottom of the window, which opens the **Add Partition on /dev/sd...** screen.
8. Select **Primary Partition**, then select **Next**. The **New Partition Size** screen displays.
9. Select **Maximum Size**, then select **Next**. The **Formatting Options and Mounting Options** screen displays.
10. Select **Format partition** and select **Ext3** from the pull-down menu under **File system**.
11. Select **Mount partition** and select `/` from the pull-down menu under **Mount Point**.
12. Select **Finish**. Partitioning of the target boot device for the operating system installation has been completed. On the **Expert Partitioner** screen, select **Accept** to accept the changes. A pop-window stating the following will be displayed:

```
Changes in disk partitioning were detected since the time the bootloader
was configured.
```

```
Do you want to proposed bootloader configuration again?
```

```
If yes, all previous bootloader configuration will be lost.
```

```
If not, you probably need to change the configuration manually.
```

Select **OK**. This pop-up window has a count-down timer that selects **OK** automatically if not interrupted with the `Stop` button. The display returns to the **Installation Settings** screen.

- c. Confirm the disk partitions.

A single `root` device and a single `swap` device, both associated with `sda` are the only partitions listed. For example:

Partitioning

Create root partition /dev/sda2 (146.92 GB) with ext3
use /dev/sda1 as swap

To make additional, site-specific changes, select the **Partitioning** section header and return to the **Expert Partitioner** screen.

23. Confirm the language for the SMW. English (US) is the default language. To change the primary language select the **Language** heading in the **Installation Settings** screen. The **Languages** window opens. Select a language from the drop-down menu. Select multiple secondary languages, if desired. Then select **Accept** at the bottom of the window.
24. On the **Installation Settings** screen, select **Install** at the bottom of the screen. The **Confirm Installation** window appears. To check or change settings, select **Back**; otherwise, select **Install** to confirm and to install the operating system.
The installation runs for approximately 30 minutes. The process automatically reboots the SMW from the hard disk, and the installation continues with system configuration.
25. The **Network Settings** window for configuring the customer network appears. Select the entry labeled **eth0 Customer Network Ethernet**. On the **Overview** tab, select edit and do the following:
 - a. Enter the IP address.
 - b. Enter the subnet mask.
 - c. Enter the short hostname and select **Next**.
 - d. Select the **Hostname/DNS** tab. In the **Hostname/DNS and Name Server Configuration** window, enter the hostname, the domain name, the name server values, and the domain names to search. Enter the hostname and domain name separately.

Host Name	Domain Name
smwhost	my.domain.com

If a fully-qualified hostname that includes the domain name is entered, the hostname is accepted but the periods are removed; for example, a hostname of `smwhost.my.domain.com` is converted to `smwhostmydomaincom`.

- e. Select the **Routing** tab. In the **Routing** window, enter the default gateway IP address. Then select **OK**.
26. The **Clock and Time Zone** window appears.
 - a. Select the appropriate time zone.
 - b. If necessary, adjust the time of day.
 - c. Verify that **Hardware Clock Set To UTC** is selected.
 - d. Select **OK**.
At this point, the system finishes booting and enters multiuser mode.
The SMW base operating system, `Cray-SMWbase11SP3`, is now installed.
27. Remove the protective covers from the Fibre Channel (FC) or SAS cables and connectors, clean the ends of the cables and connectors, and reconnect the data cables.
28. Re-seat the previously ejected SMW internal disk drives.

29. Reboot the SMW to allow the SMW to discover the drives properly.

```
smw# reboot
```

30. Use procedure [R815 SMW: Enable an Integrated Dell Remote Access Controller \(iDRAC\)](#) to enable iDRAC on an R815 SMW.

Installation of the base operating system for the R815 SMW is now complete. The system is now ready for installation and configuration of the SMW software. Go to [Install and Configure SMW Software](#) on page 30.

Install and Configure SMW Software

This procedure takes approximately 30 minutes.

1. Log on to the SMW as `crayadm`, open a terminal window, and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. If the base operating system DVD (Cray-SMWbase11SP3-) is still in the CD/DVD drive, eject it.

```
smw# eject
```

3. Place the Cray 7.2UP04 Software DVD in the drive and mount it.

```
smw# mkdir -p /media/cdrom
smw# mount /dev/cdrom /media/cdrom
```

NOTE: If problems occur while mounting the DVD after the initial Linux installation, reboot the SMW and perform this procedure again, from the beginning.

4. Copy the `SMWinstall.conf` file from `/media/cdrom` to `/home/crayadm`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
```

5. Change the permissions to make the file writable by the user only.

```
smw# chmod 644 /home/crayadm/SMWinstall.conf
```

6. Edit the `SMWinstall.conf` file to customize it for the installation site. The `SMWinstall.conf` file contains settings for:

- the system interconnection network type, e.g., Aries or Gemini
- the name of the local Network Time Protocol (NTP) servers
- configuring SMWs that have additional disks
- configuring an SMW for a Cray XE6m, Cray XK6m, or Cray XK7m mid-range system
- configuring the Cray Lightweight Log Manager (LLM)
- enabling or disabling the Linux `sar` service on the SMW

```
smw# vi /home/crayadm/SMWinstall.conf
```

a. Specify the system interconnection network type. There is no default setting.

- b. Adjust LLM settings, as needed.
By default, the `SMWinstall` program enables LLM. To send the logs from the SMW to a site loghost, adjust these settings: `LLM_siteloghost`, `LLM_sitecompatmode`, and `LLM_altrelay`. If the boot RAID controller IP addresses do not start the specified pattern, adjust `llm_raid_ip`.
- c. Leave the Linux `sar` service on the SMW enabled (default), or disable it, as needed.
- d. Leave the controller log forwarding on the SMW enabled (default), or disable it, as needed.
- e. Set the `LOGDISK`, `DBDISK`, and `PMDISK` variables.
The disk device specified by the `LOGDISK` variable must have more disk space than the device specified by the `DBDISK` variable.
1. Use the `LOGDISK` variable to specify the disk device name to be used for logging (`/var/opt/cray/log`, `/var/opt/cray/dump`, and `/var/opt/cray/debug`). The entire disk will be formatted for this use. For the appropriate disk name, see the `LOGDISK` section of the `SMWinstall.conf` file shown below.
IMPORTANT: For an SMW HA system, define this variable as if the system were a stand-alone SMW. The shared storage for logging will be configured later in the SMW HA configuration process.
 2. Use the `DBDISK` variable to specify the disk device name to be used for the SMW HSS database (`/var/lib/mysql`). The entire disk will be formatted for this use. For the appropriate disk name for rack-mount models, see the `DBDISK` section of the `SMWinstall.conf` file shown below.
IMPORTANT: For an SMW HA system, define this variable as if the system was a stand-alone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.
 3. (Optional) Use the `PMDISK` variable to specify the disk device name to be used for the PostgreSQL Database (`/var/lib/pgsql`). The entire disk will be formatted for this use. For the appropriate disk name for rack-mount models, see the `PMDISK` section of the `SMWinstall.conf` file shown below.
IMPORTANT: For an SMW HA system, do not set the `PMDISK` or `PMMOUNT` variables unless PMDB storage will **not** be shared. These variables are not used on a system that will configure shared PMDB storage as described in [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172 or [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. (The disk device for `/var/lib/pgsql` is set up later in the SMW HA configuration process.) Do not remove the comment character for `PMDISK` and `PMMOUNT` in the `SMWinstall.conf` file.

The `LOGDISK`, `DBDISK`, and `PMDISK` variables must be persistent disk device names. To use the default persistent disk device names, remove the comment character (`#`) from the relevant lines in the `SMWinstall.conf` file:

```
#For SLES 11 SP3-based rackmount R815 SMWs:
#LOGDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#LOGDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0

#LOGMOUNT=/var/opt/cray/disk/1

#For SLES 11 SP3-based rackmount R815 SMWs:
#DBDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#DBDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
```

```
#DBMOUNT=/var/lib/mysql

#For SLES 11 SP3-based rackmount R815 SMWs:
#PMDISK=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
#For SLES 11 SP3-based rackmount R630 SMWs:
#PMDISK=/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0

#PMMOUNT=/var/lib/pgsql
```

For a nonstandard configuration, modify the persistent disk device names accordingly.

- f. For Cray XE6m, Cray XK6m, or Cray XK7m systems: Set the `MCLASS` and `NumChassis` variables. These configuration settings do not apply to Cray XE6m-200, Cray XK6m-200, or Cray XK7m-200 systems.
 1. For a Cray XE6m, Cray XK6m, or Cray XK7m system, set the `MCLASS` variable to `TRUE`. For all other Cray systems, including Cray XE6m-200, Cray XK6m-200, and Cray XK7m-200 systems, `MCLASS` must retain the default setting of `FALSE`.
 2. For a Cray XE6m, Cray XK6m, or Cray XK7m system only, remove the comment character from the `NumChassis` line. The `NumChassis` variable indicates the number of chassis that are in the Cray `MCLASS` system. Choices are 1, 2, 3, 4, 6, 8, 9, 12, 15, or 18.
7. Execute the `SMWinstall` installation script, which updates the base operating system software with SMW security updates and SMW software.

```
smw# /media/cdrom/SMWinstall
```

The output of the installation script is displayed on the console. The `SMWinstall` installation script also creates log files in `/var/adm/cray/logs`.

If for any reason this script fails, it can be rerun without adverse side effects. However, rerunning this script can generate numerous error messages as the script attempts to install already-installed RPMs. Ignore these particular messages.

8. Reboot the SMW.

```
smw# reboot
```

Bootstrap Hardware Discovery for the Second SMW

For the second SMW in an SMW HA system, update the controller image. Do not perform the hardware discovery; this task was done on the first SMW.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. Update the controller boot image. The version used in the command argument for `hss_make_default_initrd` should match the version specified in the `lsb-cray-hss` line in output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Wed May 15 12:13:32 CDT 2013 on hssbld0 by bwdev
```

```
lsb-cray-hss-7.2.0-1.0702.31509.447
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.31509.447
::: Verifying base RPM list to the manifest
(additional status messages)
::: Removing unwanted files from the root

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.32944.237.
Running hssclone.
Image Clone Complete: /opt/cray/hss-images/default
Running hsspackage.
. . .(additional status messages). . .
inking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img
```

Change the Default SMW Passwords

After completing the installation, change the default SMW passwords on the SMW. The SMW contains its own `/etc/passwd` file that is separate from the password file for the rest of the system. To change the passwords on the SMW, log on to the SMW as `root` and execute the following commands:

```
crayadm@smw> su - root
smw# passwd root
smw# passwd crayadm
smw# passwd cray-vnc
smw# passwd mysql
```

For rack-mount SMWs (both R815 and R630 models), it is also necessary to change the default iDRAC password. See [Change the Default iDRAC Password](#).

Set Up the SUSE Firewall and IP Tables

The SMW software includes a firewall. The following steps enable and configure the firewall.

TIP: It is not necessary to shut down the system before performing this task.

1. Before modifying the SUSE Firewall settings, make a copy of the configuration file:

```
smw# cp -p /etc/sysconfig/SuSEfirewall12 /etc/sysconfig/SuSEfirewall12.orig
```

2. Using the `SuSEfirewall12` program and the following steps, change the IP tables rules to close off all unnecessary ports on the SMW.

```
smw# iptables -L
smw# vi /etc/sysconfig/SuSEfirewall12
```

Change the settings of these variables to the values shown:

```
FW_DEV_EXT="any eth0"

FW_DEV_INT="eth1 eth2 eth3 eth4 lo"

FW_SERVICES_EXT_UDP="161"

FW_TRUSTED_NETS="your_bootnode_ipaddress,tcp,7004 \
your_syslognode_ip,udp,514 your_sdbnode_ip,tcp,6811:6815"
```

For example:

```
smw# diff /etc/sysconfig/SuSEfirewall2.orig /etc/sysconfig/SuSEfirewall2
99c99
< FW_DEV_EXT="eth-id-00:30:48:5c:b0:ee eth0"
---
> FW_DEV_EXT="any eth0"
114c114
< FW_DEV_INT="eth-id-00:0e:0c:b4:df:64 eth-id-00:0e:0c:b4:df:65
   eth-id-00:0e:0c:b4:df:66 eth-id-00:0e:0c:b4:df:67 eth1 eth2 eth3 eth4"
---
> FW_DEV_INT="eth1 eth2 eth3 eth4 lo"
263c263
< FW_SERVICES_EXT_UDP=""
---
> FW_SERVICES_EXT_UDP="161"
394c394
< FW_TRUSTED_NETS=""
---
> FW_TRUSTED_NETS="10.3.1.254,tcp,7004 10.5.1.2,udp,514 10.5.1.2,tcp,6811:6815"
```

NOTE: 10.3.1.254 is the boot node's IP address for eth0 on the network between the boot node and the SMW. 10.5.1.2 is used for the DRBD connection on SMW HA, but is different for the first SMW and the second SMW: the first is 10.5.1.2 (active) and the second is 10.5.1.3 (passive).

3. Invoke the modified configuration.

```
smw# /etc/init.d/SuSEfirewall2_init start
smw# /etc/init.d/SuSEfirewall2_setup start
```

4. Execute the following commands to start the firewall at boot time.

```
smw# chkconfig SuSEfirewall2_init on
smw# chkconfig SuSEfirewall2_setup on
```

5. Verify the changes to the iptables.

```
smw# iptables -nvL
```

SSH access is one of the protocols permitted through the firewall from the external network to the SMW. For information about how to use Virtual Network Computing (VNC) through an SSH tunnel, see [Enable Remote Access to the SMW using VNC](#).

Configure the Simple Event Correlator (SEC)

The System Management Workstation (SMW) 7.2.UP04 release includes the Open Source simple event correlator (SEC) package, `sec-2.7.0`, and an SEC support package, `cray-sec-version`. The SEC support package contains control scripts to manage the starting and stopping of SEC around a Cray mainframe boot session, in addition to other utilities.

To use the Cray SEC, see *Configure SEC Software (S-2542)* for configuration procedures.

Finish Installing SMW Software on the Second SMW

Prerequisites

Before using this procedure, ensure that the operating system and Cray SMW software is correctly installed on the first SMW.

Use these HA-specific steps to finish the SMW software installation on the second SMW.

1. If you changed the default root password on the SMW, you must also change the default iDRAC password to the same password. For more information, see [Passwords For an SMW HA System](#) on page 18.

IMPORTANT: Both SMWs must have exactly the same password for the root and iDRAC accounts.

2. Configure email on the SMW. The SMW HA system uses email for failover notification. For information about configuring email on an SMW, see http://www.postfix.org/BASIC_CONFIGURATION_README.html.
3. If there are local changes to `/etc/hosts` on `smw1`, manually copy `/etc/hosts` to `/etc/hosts` on `smw2`. The customized entries must be above the first section of "XT Cabinet x - y".

```
smw2:~ # cp /etc/hosts /etc/hosts.sav
smw2:~ # scp smw1:/etc/hosts /etc/hosts
```

- a. Edit the `/etc/hosts` file on `smw2` to change the line `smw1-ip smw1 smw1` to `smw2-ip smw2 smw2`
4. Ensure that the boot RAID is connected before you continue to the next procedure.

Install CLE Software on the Second SMW

After installing the the SMW HA software on the second SMW, use this procedure to install the CLE software on the SMW.

1. Copy the CLE install directory, `/home/crayadm/install.xtre1`, from the first SMW to a local directory on the second SMW (such as `/tmp`), where `xtre1` is the site-determined name specific to the release being installed. Do not use `/home/crayadm` on the second SMW, because that would create local differences for this shared directory.
2. As `root`, execute the following command to install the Cray CLE software on the second SMW. Include the `-X Aries` option to prevent an `xtdiscover` error on the second SMW.

```
smw2:~ # /tmp/install.xtre1/CRAYCLEinstall.sh \
-m /tmp/install.xtre1 -X Aries -v -i -w
```

This example shows `/tmp` as the location of the CLE install directory on the second SMW.

3. At the prompt `'Do you wish to continue?'`, type `y` and press Enter.

The output of the installation script displays on the console. If this script fails, restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. Do not be concerned about these messages.

NOTE: If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

4. Configure firewall and route file changes on the second SMW.
 - a. Copy the file `/var/opt/cray/install/networking_configuration-p0.json` from the first SMW to the second SMW.
 - b. Execute the following command on `smw2`.

```
smw2:~ # /opt/cray/keystone-cle-config/default/bin/cray_configure_networking.py \
--enable --smw --partition p0 \
--networkjson /var/opt/cray/install/networking_configuration-p0.json \
--servicesjson /opt/cray/keystone-cle-config/default/etc/openstack_services.json \
--nopretend
```

Install SMW HA Software

After the SMW and CLE software has been installed, verify the system configuration, then install the SMW HA release package on both SMWs.

Verify the SMW and CLE Configuration

Before installing the SMW HA release package, ensure that both SMWs are running the same SMW and CLE software. In addition, the boot RAID must be set up for SMW HA before configuring the SMW HA software release package.

1. Log in as `root` to both SMWs.
2. Ensure that both SMWs have the same root password, and that the iDRAC password on each SMW is the same as the root password. For more information, see [Passwords For an SMW HA System](#) on page 18.
3. Verify that both SMWs are running the same SMW release.

```
smw1:~ # cat /opt/cray/hss/default/etc/smw-release
7.2.UP04
```

```
smw2:~ # cat /opt/cray/hss/default/etc/smw-release
7.2.UP04
```

4. Verify that both SMWs are running the same CLE release.

```
smw1:~ # cat /etc/opt/cray/release/CLEinfo
CLERELEASE=5.2.UP04
INSTALLERVERSION=d01
LUSTRE=yes
NETWORK=ari
XTRELEASE=5.2.52
```

```
smw2:~ # cat /etc/opt/cray/release/CLEinfo
CLERELEASE=5.2.UP04
INSTALLERVERSION=d01
```

```
LUSTRE=yes
NETWORK=ari
XTRELEASE=5.2.52
```

5. Verify that both SMWs are running the same operating system release.

```
smw1:~ # cat /etc/SuSE-release
USE Linux Enterprise Server 11 (x86_64)
VERSION = 11
PATCHLEVEL = 3
```

```
smw2:~ # cat /etc/SuSE-release
USE Linux Enterprise Server 11 (x86_64)
VERSION = 11
PATCHLEVEL = 3
```

6. Ensure that kernel 3.0.101-0.461 (which was provided in FN-6029), or a later kernel update has been applied. There is a SLES kernel dependency on the `ocfs2-kmp-default` RPM package that will prevent some SLES HA RPMs from being installed unless this kernel update has been applied.
7. Check that the boot RAID has free LUNS with sufficient space for the shared directories. For more information, see [Configure the Boot RAID for SMW HA](#) on page 47.
8. Verify the iDRAC configuration.
 - a. Log in as `root` on both SMWs.
 - b. Check the iDRAC configuration on each SMW with the following commands:

NOTE: Replace `smw#-iDRAC-IP-addr` with the SMW's iDRAC IP address.

```
smw1:~ # ipmitool -U root -I lanplus -H smw1-iDRAC-IP-addr -a chassis \
power status
Password:
Chassis Power is on
```

```
smw2:~ # ipmitool -U root -I lanplus -H smw2-iDRAC-IP-addr -a chassis \
power status
Password:
Chassis Power is on
```

If the iDRAC is configured correctly, these commands return the output `Chassis Power is on` (or `off`). If either of these commands fails, the iDRAC is not configured correctly.

9. Check that both SMWs and iDRACs appear in the DNS.

NOTE: In the following commands, specify the actual SMW and iDRAC host names.

```
smw1:~ # ping smw2
smw1:~ # ping smw2-iDRAC-hostname

smw2:~ # ping smw1
smw2:~ # ping smw1-iDRAC-hostname
```

If the iDRACs are not in the DNS (or if the DNS is not available), you can add the iDRAC entries to `/etc/hosts` on both SMWs.

Install the SMW HA Release Package on Both SMWs

Start the SMW HA installation on the first SMW, which was completely installed and configured with the SMW and CLE software. This SMW will initially be the active SMW when the system is fully configured. The second SMW will initially be the passive SMW.

NOTE: The examples in this procedure show the host name `smw1` for the first SMW and the host name `smw2` for the second SMW.

1. Log on to the first SMW (`smw1`) as `root`.
2. Mount the Cray SMW HA release media on the SMW.
 - a. If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw1:~ # mount /dev/cdrom /media/cdrom
```

- b. If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

NOTE:

The ISO file name depends on the release number, and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For `path`, substitute the actual path to the ISO on the system.

IMPORTANT: The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

```
smw1:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

3. Navigate to the `/media/cdrom` directory and execute the `SMWHAinstall` script to install the Cray SMW HA release software on the SMW.

```
smw1:~ # cd /media/cdrom
smw1:~ # ./SMWHAinstall -v
```

4. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.
5. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw1:~ # cd
smw1:~ # umount /media/cdrom
smw1:~ # eject
```

6. Reboot the first SMW.
7. Log on to the second SMW (`smw2`) as `root` to repeat the installation on the other SMW.
8. Mount the Cray SMW HA release media on the SMW.

- a. If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw1:~ # mount /dev/cdrom /media/cdrom
```

- b. If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

NOTE:

The ISO file name depends on the release number, and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For `path`, substitute the actual path to the ISO on the system.

IMPORTANT: The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

```
smw1:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

9. Navigate to the `/media/cdrom` directory and execute the `SMWHAinstall` script to install the Cray SMW HA release software on the SMW.

```
smw2:~ # cd /media/cdrom
smw2:~ # ./SMWHAinstall -v
```

10. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.

11. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw2:~ # cd
smw2:~ # umount /media/cdrom
smw2:~ # eject
```

12. Reboot the second SMW.

Configure the Cluster

After the operating system, SMW, CLE, and HA software has been installed on both SMWs, use the following procedures to configure the required cluster settings, boot image synchronization, failover notification, and the power management database (PMDB) on the boot RAID.

Configure Required Cluster Settings

Prerequisites

Before beginning this procedure:

- Gather the necessary information. This procedure requires site-specific configuration information from [Minimum Boot RAID LUN Sizes for SMW Failover](#) on page 15, [Fixed IP Addresses for an SMW HA system](#) on page 16, and [Site-dependent Configuration Values for an SMW HA System](#), and [Boot RAID LUNs for the Shared Directories](#), including virtual and actual host names, IP addresses, and disk names of the shared file systems on the boot RAID.
- To retain typescript sessions for this install procedure, Cray recommends starting a typescript for each SMW on a local workstation:

```
workstation> script -af my_output_file
Script started, file is my_output_file
workstation> ssh root@smw1
```

Alternatively, create a typescript session in the root home directory and restart the session after the system reboots.

Use the following procedure to configure the required SMW HA cluster settings. During this procedure, the first SMW (`smw1`) becomes the active SMW. The second (`smw2`) becomes the passive SMW.

1. Log into the first SMW (`smw1`) as `root`. Log in directly as `root`; do not use `su` from a different account.

```
workstation> ssh -X root@smw1
```

2. In a separate terminal session, log into the other SMW (`smw2`) as `root`. Log in directly as `root`; do not use `su` from a different account.

```
workstation> ssh -X root@smw2
```

3. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, shut down the VNC server.

```
smw1:~ # /etc/init.d/vnc stop
```

4. Update the `ifcfg` files for `eth2` and `eth4` on `smw1`.

- a. Edit the `/etc/sysconfig/network/ifcfg-eth2` file.

```
smw1:~ # vi /etc/sysconfig/network/ifcfg-eth2
```

- b. Change the `NAME` value from `'eth2 Reserved'` to `'eth2 SMW HA Heartbeat Network 1'`.
- c. Verify your changes. The file must have the following contents:

```
BOOTPROTO='static'
IPADDR='10.2.1.1/16'
NAME='eth2 SMW HA Heartbeat Network 1'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

Do not change the `IPADDR` value. The HA configuration process determines the actual values automatically.

- d. Save your changes and exit the editor.
- e. Copy the `ifcfg-eth2` file to `ifcfg-eth4`.

```
smw1:~ # cp /etc/sysconfig/network/ifcfg-eth2 /etc/sysconfig/network/ifcfg-eth4
```

- f. Edit the `/etc/sysconfig/network/ifcfg-eth4` file.

```
smw1:~ # vi /etc/sysconfig/network/ifcfg-eth4
```

- g. Change the `IPADDR` value from `'10.2.1.1/16'` to `'10.4.1.1/16'`.

```
IPADDR='10.4.1.1/16'
```

Use the specified `IPADDR` value on both SMWs. The HA configuration process determines the actual values automatically.

- h. Change the `NAME` value to `'eth4 SMW HA Heartbeat Network 2'`.

- i. Verify the changes. The file must have the following contents:

```
BOOTPROTO='static'
IPADDR='10.4.1.1/16'
NAME='eth4 SMW HA Heartbeat Network 2'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- j. Save your changes and exit the editor.

5. Repeat the previous step on `smw2` to update the `ifcfg` files for `eth2` and `eth4` on `smw2`.

- a. Edit the `/etc/sysconfig/network/ifcfg-eth2` file.

```
smw2:~ # vi /etc/sysconfig/network/ifcfg-eth2
```

- b. Change the `NAME` value from `'eth2 Reserved'` to `'eth2 SMW HA Heartbeat Network 1'`.

- c. Verify your changes. The file must have the following contents:

```
BOOTPROTO='static'
IPADDR='10.2.1.1/16'
NAME='eth2 SMW HA Heartbeat Network 1'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

Do not change the `IPADDR` value. The HA configuration process determines the actual values automatically.

- d. Save your changes and exit the editor.

- e. Copy the `ifcfg-eth2` file to `ifcfg-eth4`.

```
smw2:~ # cp /etc/sysconfig/network/ifcfg-eth2 /etc/sysconfig/network/ifcfg-eth4
```

- f. Edit the `/etc/sysconfig/network/ifcfg-eth4` file.

```
smw2:~ # vi /etc/sysconfig/network/ifcfg-eth4
```

- g. Change the `IPADDR` value from `'10.2.1.1/16'` to `'10.4.1.1/16'`.

```
IPADDR='10.4.1.1/16'
```

Use the specified `IPADDR` value on both SMWs. The HA configuration process determines the actual values automatically.

- h. Change the `NAME` value to `'eth4 SMW HA Heartbeat Network 2'`.
- i. Verify the changes. The file must have the following contents:

```
BOOTPROTO='static'
IPADDR='10.4.1.1/16'
NAME='eth4 SMW HA Heartbeat Network 2'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- j. Save your changes and exit the editor.

6. Update the cluster IP addresses.

- a. On `smw1`, execute the following command with `0` as the first argument.

NOTE: In this command, replace `smw1` with the host name of the first SMW, and replace `smw2` with the host name of the second SMW.

```
smw1:~ # /opt/cray/ha-smw/default/hainst/update_addresses 0 smw1 smw2
```

- b. On `smw2`, execute this command with `1` as the first argument:

NOTE: In this command, replace `smw1` with the host name of the first SMW, and replace `smw2` with the host name of the second SMW.

```
smw2:~ # /opt/cray/ha-smw/default/hainst/update_addresses 1 smw1 smw2
```

7. Initialize `smw1` as the active SMW.

- a. Execute the `sleha-init` command on `smw1`.

```
smw1:~ # sleha-init
```

You can safely ignore error messages about file `/etc/corosync/corosync.conf`. This file will be created after `sleha-init` completes.

```
awk: cmd. line:1: fatal: cannot open file `/etc/corosync/corosync.conf'
for reading
(No such file or directory)"
```

- b. As `sleha-init` runs, it prompts you for required information. Answer the following questions to configure the cluster.

```
Network address to bind to (e.g.:192.168.1.0): 10.2.1.0
Multicast address (e.g.:239.x.x.x): 226.0.0.1
Multicast port [5405]: 1694
Configure SBD:
.
.
.
Do you wish to use SBD? [y/N]: N
WARNING: Not configuring SBD - STONITH will be disabled.
.
.
```

```
Done (log saved to /var/log/sleha-bootstrap.log)
```

If SMW HA has been configured before and you wish to rerun `sleha-init`, you will also be prompted to overwrite the existing configuration. In this case, answer the prompt `* - overwrite [y/N]?` with **y**.

c. Wait for `sleha-init` to finish (normally, about 1 or 2 minutes).

8. Join `smw2` as the passive SMW.

a. Execute the `sleha-join` command on `smw2`.

```
smw2:~ # sleha-join
```

b. Answer the following questions to join the passive SMW to the cluster. When asked for the password, use the `root` password for the SMWs.

NOTE: In this command, replace `smw1` with the host name of the first SMW.

```
IP address or hostname of existing node (active SMW): smw1
```

```
Password: root-password-for-SMWs
```

9. Check the cluster status to verify that both `smw1` and `smw2` are online.

```
smw1 # crm_mon -r1 | grep Online
Online: [ smw1 smw2 ]
```

The `crm_mon` command displays the SMW host names in alphanumeric order; the first SMW shown is not necessarily the active SMW.

10. Configure `eth4` as the redundant heartbeat channel on `smw1`.

a. Execute `yast2` to open the YaST2 Control Center.

```
smw1:~ # yast2 cluster
```

For the GUI version of YaST, either execute this command on the SMW console or connect via an `ssh` connection with X11 port forwarding (for example, `ssh -X smw1`).

The cluster wizard starts and opens the cluster configuration window.

b. In the left panel, ensure that Communication Channels is selected.

c. In the right panel, check the Redundant Channel check box, then enter the following information to configure `eth4` as the redundant channel:

- Bind Network Address: **10.4.1.0**
- Multicast Address: **225.0.0.1**
- Multicast Port: **1694**

IMPORTANT: Be careful to start the multicast address with 225, not 255. An incorrect multicast address will prevent the cluster from starting.

d. Click the Finish button.

e. Close the main YaST2 window to exit `yast2`.

11. Configure `eth4` as the redundant heartbeat channel on `smw2`.

- a. Execute `yast2` to open the YaST2 Control Center.

```
smw2:~ # yast2 cluster
```

For the GUI version of YaST, either execute this command on the SMW console or connect via an `ssh` connection with X11 port forwarding (for example, `ssh -X smw2`).

The cluster wizard starts and opens the cluster configuration window.

- b. In the left panel, ensure that **Communication Channels** is selected.
- c. In the right panel, check the Redundant Channel check box, then enter the following information to configure `eth4` as the redundant channel:

- Bind Network Address: `10.4.1.0`
- Multicast Address: `225.0.0.1`
- Multicast Port: `1694`

IMPORTANT: Be careful to start the multicast address with **225**, not 255. An incorrect multicast address will prevent the cluster from starting.

- d. Click the Finish button.
- e. Close the main YaST2 window to exit `yast2`.

12. On `smw1`, synchronize the passive SMW.

```
smw1:~ # csync2 -xv
```

13. Reset the login environment on both SMWs by logging out, then logging back in as `root`.

You must log in to the actual (not virtual) SMW as `root`. Do not use `su` from a different account.

```
smw1:~ # exit
workstation> ssh root@smw1
```

```
smw2:~ # exit
workstation> ssh root@smw2
```

14. Edit the SMW HA configuration file, `/opt/cray/ha-smw/default/hainst/smwha_args`, to configure the site-specific settings. Replace the following default values with the actual values for the site.

NOTE: For required host names and IP addresses, see [Site-dependent Configuration Values for an SMW HA System](#). For the persistent device names for the shared directories on the boot RAID, see [Boot RAID LUNs for the Shared Directories](#).

```
smw1:~ # vi /opt/cray/ha-smw/default/hainst/smwha_args

--virtual_hostname
cray-smw
--virtual_ip
172.31.73.165
--log_disk_name
/dev/disk/by-id/scsi-360080e500023bff6000006b1515d9bc9
--db_disk_name
/dev/disk/by-id/scsi-360080e500023bff6000006b3515d9bdf
```

```

--home_disk_name
/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9c01
--drac_ip_active
172.31.73.142
--drac_ip_passive
172.31.73.77
--passive_smw_hostname
cray-smw2
--verbose

```

15. Execute the following command on `smw1` to ensure that `/home/crayadm/.gvfs` is not mounted.

```
smw1:~ # df -a | grep /home/crayadm/.gvfs && umount -f /home/crayadm/.gvfs
```

16. Ensure that nothing is mounted on `/mnt`. The `SMWHAconfig` script uses `/mnt` to set up the shared storage.

```
smw1:~ # df -a | grep mnt
smw1:~ #
```

17. Configure the SMW HA cluster on the active SMW.

- a. Change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

- b. Execute `SMWHAconfig` on `smw1` only, using the modified configuration file as an argument (prefaced by the `@` character). If necessary, answer a prompt or perform the specified action to complete the `ssh` connection.

```

smw1:~ # ./SMWHAconfig @smwha_args
2014-08-22 11:1:56,156: INFO      cdir was created
2014-08-22 11:31:56,361: INFO
*****Starting of HA software
installation*****

2014-08-22 11:31:56,361: INFO      cluster virtual IP = 172.31.73.165
2014-08-22 11:31:56,361: INFO      log disk (/var/opt/cray/disk/1) = /dev/
disk/by-id/scsi-360080e500023bff6000006b1515d9bc9
2014-08-22 11:31:56,361: INFO      db disk (/var/lib/mysql)= /dev/disk/by-id/
scsi-360080e500023bff6000006b3515d9bdf
2014-08-22 11:31:56,362: INFO      home disk (/home)= /dev/disk/by-id/
scsi-360080e500023bff6000006b5515d9c01
2014-08-22 11:31:56,362: INFO      verbose mode =
.
.
.

```

- c. When the `SMWHAconfig` command prompts for a password so that it can configure the SMW HA cluster and the iDRAC; enter the `root` password for the SMW.
- d. Wait while `SMWHAconfig` automatically loads the HA cluster configuration settings. `SMWHAconfig` deletes any existing data in the shared directories on the boot RAID. For an initial installation, existing data is not reused.
- e. If necessary, examine the log file. `SMWHAconfig` creates a log file in `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out`. You can safely ignore the warning message that the CIB has no configuration element, if this message appears in the `SMWHAconfig` output.

18. Synchronize the `ssh` host keys. This step makes both SMWs appear to have the same `ssh` host identity when someone connects to the virtual SMW host name or IP address.

NOTE: In the following commands, replace `smw2` with the host name of the passive SMW.

- a. On `smw1`, copy the `ssh` host keys to `smw2`.

```
smw1:~ # scp -p /etc/ssh/ssh_host_*key* root@smw2:/etc/ssh
```

- b. On `smw2`, restart the `ssh` daemon.

```
smw2:~ # /etc/init.d/sshd restart
```

- c. On `smw1`, refresh the `ssh` host keys.

```
smw1:~ # ssh-keygen -R smw2
```

- d. On `smw1`, verify that passwordless `ssh` is still functional to the other SMW.

```
smw1:~ # ssh smw2
```

After running `ssh`, you might need to answer the prompt or perform the specified action to complete the connection.

19. Use the `passwd` command to set the password for user `hacluster` on both SMWs. Follow the prompts. Note that the password **must** be the same as the SMW root password.

```
smw1# passwd hacluster
smw2# passwd hacluster
```

20. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

21. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

22. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```

Online: [ smw1 smw2 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                    Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                     Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql   (lsb:postgresql):             Started smw1
  mysqld       (ocf::heartbeat:mysql):     Started smw1

```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```

smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

23. If necessary, restart the VNC server.

Configure Boot Image Synchronization

To ensure that boot images are properly synchronized for the SMW HA system, you must set up the boot image directory. The procedure depends on whether boot images are specified in the `/etc/sysset.conf` configuration file as files in a `/bootimagedir` directory or as images on a raw device (such as `/raw0`).

NOTE: Cray recommends storing boot images as files in a `/bootimagedir` directory. If the boot image directory is a raw device (such as `/raw0`), a change is required on `smw2` to allow the SMW HA system to synchronize the boot images. For more information, see [Create Boot Images](#) on page 99.

Configure Boot Image Synchronization for Images in a `/bootimagedir` directory

If boot images are specified as files in a `/bootimagedir` directory, use this procedure to configure boot image synchronization.

1. Log in as root on the first SMW (`smw1`).

2. Edit the file `/etc/csync2/csync2_cray.cfg`.

```
smw1:~ # vi /etc/csync2/csync2_cray.cfg
```

3. In the `group user_group` section, add an entry for `/bootimagedir` using the following format:

NOTE: Replace `bootimagedir` with the name of the boot image directory.

```
include /bootimagedir/*;
```

4. Save the changes and exit the editor.

5. If necessary, create the `/bootimagedir` directory on `smw2`. The boot image directory must exist on both SMWs.

```
smw2:~ # mkdir -p /bootimagedir
```

6. Copy the boot images from `smw1` to `smw2` to initialize boot image synchronization. This manual copy operation speeds up future synchronization. Execute the following command for each boot image in the boot image directory.

```
smw1:~ # scp -pr /bootimagedir/* smw2:/bootimagedir/
```

7. (Optional) Copy the `xt-images` file from `smw1` to `smw2`. Although this step is not required for normal operations, having a copy of this file on the second SMW is necessary for building CLE boot images on the second SMW, and for certain SEC error-reporting purposes.

```
smw1:~ # scp -pr /opt/xt-images smw2:/opt
```

Configure Boot Image Synchronization for Images on a Raw Device

If the boot image directory is specified as a raw device (for example `/raw0`), use this procedure to configure boot image synchronization.

IMPORTANT: The boot image directory must exist on both SMWs.

1. Log in as `root` on the second SMW (`smw2`).
2. Create a symbolic link from the physical device name to the raw device.

NOTE: In the following command, replace `/dev/disk/by-id/xxxx` with the persistent device name for the actual device; replace `/rawdevice` with the raw device name.

```
smw2:~ # ln -s /dev/disk/by-id/xxxx /rawdevice
```

Configure Failover Notification

Prerequisites

Failover notification requires email to be configured on both SMWs. For information about configuring email, see http://www.postfix.org/BASIC_CONFIGURATION_README.html.

The SMW HA software includes a `Notification` resource that automatically sends email when a failover occurs.

You can configure failover notification either during initial installation or after the HA system is installed and running.

1. Execute the `crm resource` command.

```
smw1:~ # crm resource param Notification set email address@thedomain.com
```

NOTE: Only one email address is allowed. To send notifications to multiple addresses, you can create a group email alias that includes these email addresses.

2. Verify the setting.

```
smw1:~ # crm resource param Notification show email address@thedomain.com
```

If a failover occurs, the `Notification` resource sends several messages that are similar to the following examples.

```
From: root [mailto:root@smw.none]
Sent: Thursday, June 06, 2013 9:20 PM
To: Cray Cluster Administrator
Subject: ***Alert*** A Failover may have occurred. Please investigate!
Migrating resource
away at Thu Jun 6 21:20:25 CDT 2013 from smw1
```

```
***Alert*** A Failover may have occurred. Please investigate!
Migrating resource away
at Thu Jun 6 21:20:25 CDT 2013 from smw1
```

```
Command line was:
/usr/lib/ocf/resource.d//heartbeat/MailTo stop
From: root [mailto:root@smw.none]
Sent: Thursday, June 06, 2013 9:20 PM
To: Cray Cluster Administrator
Subject: ***Alert*** A Failover may have occurred. Please investigate!
Takeover in progress
```

```
at Thu Jun 6 21:20:25 CDT 2013 on smw2
```

```
***Alert*** A Failover may have occurred. Please investigate!
Takeover in progress
```

```
at Thu Jun 6 21:20:25 CDT 2013 on smw2
```

```
Command line was:
```

```
/usr/lib/ocf/resource.d//heartbeat/MailTo start
```

Configure PMDB Storage

Choose one of these options to configure shared storage for the Power Management Database (PMDB).

- [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172. Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.
- [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. Shared storage: A logical disk, configured as a LUN (Logical Unit) or logical volume on the boot RAID. The boot RAID must have sufficient space for `/var/lib/pgsql`.

Cray strongly recommends using either mirrored storage (preferred) or shared storage. An unshared PMDB is split across both SMWs; data collected before an SMW failover will be lost or not easily accessible after failover. For more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9.

Configure Mirrored Storage with DRBD for the PMDB

Prerequisites

IMPORTANT:

If mirrored storage becomes available after the PMDB has been configured for shared storage, use the procedure [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322 instead of this procedure.

Before beginning this procedure:

- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- Plan sufficient time for this procedure. Transferring the Power Management Database (PMDB) to a 1 TB disk requires about 10 hours. The SMW HA cluster should be in maintenance mode until the synchronization operation completes. The Cray system (compute and service nodes) can remain up and can run jobs during this period.
- Check `/etc/fstab` to ensure that there is no entry for `phy3`.
- If upgrading or updating the SMW HA system, ensure that the following RPMs are installed on both SMWs and that the version number is 8.4.4 or higher:

```
drbd-bash-completion-8.4.4-0.22.9
drbd-kmp-default-8.4.4_3.0.101_0.15-0.22.7
drbd-udev-8.4.4-0.22.9
drbd-utils-8.4.4-0.22.9
```

```
drbd-pacemaker-8.4.4-0.22.9
drbd-xen-8.4.4-0.22.9
drbd-8.4.4-0.22.9
```

If necessary, install or update any missing RPMs with "zypper install drbd".

Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.

This procedure configures the network for DRBD, configures the DRBD disks, and transfers the PMDB data from local disk to the mirrored DRBD disks.

1. Add `eth5` to the network files.

- a. Log in as root on the first SMW (*smw1*).

```
workstation> ssh root@smw1
```

- b. On *smw1*, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.2/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- c. In a separate terminal session, log in as root on the other SMW (*smw2*).

```
workstation> ssh root@smw2
```

- d. On *smw2*, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.3/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

2. Reinitialize the `eth5` interface on both SMWs.

```
smw1:~# ifdown eth5; sleep 1; ifup eth5
```

```
smw2:~# ifdown eth5; sleep 1; ifup eth5
```

3. Verify the IP addresses from *smw1*.

```
smw1:~# ping -c3 10.5.1.3
```

4. Configure the firewall to allow `eth5` as an internal connection on both SMWs.

- a. Edit the file `/etc/sysconfig/SuSEfirewall2` on both *smw1* and *smw2*.

- b. Locate the line containing the `FW_DEV_INT` variable.
- c. If necessary, add `eth5` to the end of the `FW_DEV_INT` line.

```
FW_DEV_INT="eth1 eth2 eth3 eth4 eth5 lo"
```

- d. Save your changes and exit the editor on both SMWs.
5. Reinitialize the IP tables by executing the `/sbin/SuSEfirewall12` command on both SMWs.

```
smw1:~# /sbin/SuSEfirewall12
```

```
smw2:~# /sbin/SuSEfirewall12
```

6. On the active SMW only, add the new DRDB disk to the SMW HA configuration.

NOTE: The following examples assume that `smw1` is the active SMW.

- a. Verify that the device exists on both SMWs.
For Dell R-630 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

For Dell R815 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

- b. Determine if the dedicated disk for the PMDB must be formatted. In this procedure, this disk is referred to as `PMDISK`.

NOTE: If the `PMDISK` is already correctly formatted, skip to step [6.f](#) on page 175.

This procedure assumes that a disk drive is available for use as a dedicated drive for the PMDB. The drive should be physically located within the rack-mount SMW at slot 4. The drive should be of the specification 1 TB 7.2K RPM SATA 3Gbps 2.5in HotPlug Hard Drive 342-1998, per the SMW Bill of Materials. On a Dell PowerEdge R815 the device for `PMDISK`

is `/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0` On a Dell PowerEdge R630 the device for `PMDISK` is `/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0`.

- c. Verify that the `PMDISK` is inserted into the SMW by entering the correct device name. This example is for a Dell R815.

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
```

```
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
255 heads, 63 sectors/track, 121601 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1
```

Device	Boot	Start	End	Blocks	Id	System
--------	------	-------	-----	--------	----	--------

- d. Create a new primary partition for the PMDISK, and write it to the partition table. If there are any existing partitions on this disk, manually delete them first.

```
smw:#fdisk \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)p
Partition number (1-4, default 1): 1
First sector (2048-1953525167, default 2048): [press return]
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-1953525167, default 1953525167): [press
return]
Using default value 1953525167
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

- e. Verify that the partition has been created. This should be device /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
81 heads, 63 sectors/track, 382818 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1
```

	Device	Boot	Start	End	Blocks	Id	System
	/dev/disk/by-path/. . .-lun-0-part1		2048	1953525167	976761560	83	Linux

- f. Navigate to the directory containing the SMWHAconfig command.

```
smw1:~# cd /opt/cray/ha-smw/default/hainst
```

- g. Execute SMWHAconfig to add the DRBD disk. For *disk-device*, specify the disk ID of the disk backing the DRBD disk, using either the by-name or by-path format for the device name. On a rack-mount SMW (either Dell R815 or R630), the DRBD disk is a partition on the disk in slot 4. On a Dell 815 this is /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1. On a Dell 630 it is /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1

```
smw1:~# ./SMWHAconfig --add_disk=pm-fs --device=/dev/drbd_r0 --directory=/var/lib/pgsql \
--pm_disk_name=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

7. Reboot the active SMW (*smw1*) and wait for it to boot completely.

8. Reboot the other SMW (*smw2*) and wait for it to boot completely.
9. Correct the permissions for the `/var/lib/pgsql` file on the active SMW.

```
smw1:~# chown postgres:postgres /var/lib/pgsql
smw1:~# chmod 750 /var/lib/pgsql
```

10. Put the SMW HA cluster into maintenance mode while waiting for the DRBD sync operation to complete. When *smw1* and *smw2* rejoin the cluster after rebooting, the primary DRBD disk (in *smw1*) synchronizes data to the secondary disk (in *smw2*). DRBD operates at the device level to synchronize the entire contents of the PMDB disk. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. The Cray system (service and compute nodes) can be booted and can run jobs during this period.

IMPORTANT:

Cray strongly recommends putting the SMW HA cluster into maintenance mode to prevent any failover during the sync operation. If a failover were to occur during this period, the newly-active SMW could have an incomplete copy of PMDB data.

- a. Put the SMW HA cluster into maintenance mode on *smw1*.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- b. Check the status of the DRBD sync operation with either `rcdrbd status` or `cat /proc/drbd`. The `rcdrbd` output is easier to read, but `/proc/drbd` contains more status information and includes an estimate of time to completion.

```
smw1:~# rcdrbd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res cs          ro          ds          p
mounted          fstype
0:r0 SyncSource Primary/Secondary UpToDate/Inconsistent C /var/lib/
pgsql ext3
... sync'ed:      72.7%          (252512/922140)M
```

```
smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2
ua:0 ap:0 ep:1 wo:f oos:260636656
[=====>.....] sync'ed: 72.4% (254524/922140)M
finish: 2:21:07 speed: 30,768 (29,720) K/sec
```

For an explanation of the status information in `/proc/drbd`, see the DRDB User's Guide at [linbit.com: http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd](http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd).

11. When the DRBD sync operation finishes, bring the HA cluster out of maintenance mode on *smw1*.

```
smw1:~# crm configure property maintenance-mode=false 2> /dev/null
```

12. Examine the output of `crm status` to ensure that the `ip_drbd_pgsql` is started on *smw1* and that the `Masters` and `Slaves` entries for `ms_drbd_pgsql` display the SMW host names (*smw1* and *smw2*).

```
smw1:~# crm status
Last updated: Thu Jan 22 18:40:21 2015
Last change: Thu Jan 22 11:51:36 2015 by hacluster via crmd on smw1
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.11-3ca8c3b
2 Nodes configured, 2 expected votes
23 Resources configured
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
.
.
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysqld       (ocf::heartbeat:mysql):        Started smw1
ip_drbd_pgsq  (ocf::heartbeat:IPaddr2):      Started smw1
Master/Slave Set: ms_drbd_pgsq [drbd_pgsq]
  Masters: [ smw1 ]
  Slaves: [ smw2 ]
```

Configure Shared Storage on the Boot RAID for the PMDB

Prerequisites

The SMW HA system can be configured to store the Power Management Database (PMDB) on shared storage, a logical disk configured as a LUN (Logical Unit) or logical volume on the boot RAID.

IMPORTANT: Cray strongly recommends using mirrored storage, if available, for the PMDB; for more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9. To move the PMDB from shared storage to mirrored storage, see [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322.

Before beginning this procedure:

- Ensure that the boot RAID contains a LUN for the PMDB with sufficient space for the data. Use the following command to check the size of `/var/lib/pgsql` on the local disk:

```
smw1:~ # du -hs /var/lib/pgsql
```

- Check that the boot RAID is connected.
- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- To capture typescript output from this procedure, do not use a typescript session running directly on the SMW. To save the output of this procedure, use the `script` command to start the typescript session on your local workstation before logging into the SMW, as in this example:

```
workstation> script -af my_output_file
Script started, file is my_output_file
workstation> ssh crayadm@smw1
```

Use this procedure to configure the RAID disk and transfer the power management data base (PMDB) to the power management disk on the shared boot RAID.

1. Shut down the Cray system by typing the following command as `crayadm` on the active SMW (`smw1`).

```
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
```

2. Log into the active SMW as `root`, either at the console or by using the actual (not virtual) host name.

IMPORTANT: You must log in directly as `root`. Do not use `su` from a different SMW account such as `crayadm`.

3. Change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

4. Use the `SMWHAconfig` command to move the PMDB and configure the required HA resources. In the following command, replace `scsi-xxxxxxxx` with the persistent device name for the PMDB directory on the boot RAID.

```
smw1:~ # ./SMWHAconfig --add_disk=pm-fs \
--device=/dev/disk/by-id/scsi-xxxxxxxx --directory=/var/lib/pgsql
```

This command mounts the PMDB directory (`/var/lib/pgsql`) to the boot RAID, copies the PMDB data, and configures the HA resources `pm-fs` and `postgresqld`.

5. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

6. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

7. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```

Online: [ smw1 smw2 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                    Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):            Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql   (lsb:postgresql):        Started smw1
  mysqld       (ocf::heartbeat:mysql):    Started smw1

```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```

smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

8. Verify that the Power Management Database is on the boot RAID and that the required PMDB resources are running.

- a. Examine the log file `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out` to verify that the Power Management Database disk appears in the Cluster RAID Disks section (at the end of the file), as in this example.

```
----- Cluster RAID Disks -----
07-07 20:47 INFO   MySQL Database disk = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   Log disk             = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   /home disk          = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   PM database disk    = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   ***** Ending of HA software add_disk *****
```

- b. Ensure that the power management file system is mounted by checking for `/var/lib/pgsql` in the output of the `df` command.

```
smw1:~ # df
Filesystem      1K-blocks      Used Available Use% Mounted on
/dev/sda2       120811676    82225412  32449332  72% /
udev            16433608         756  16432852   1% /dev
tmpfs           16433608         37560  16396048   1% /dev/shm
/dev/sdo        483807768   197536596  261695172  44% /var/opt/cray/disk/1
/dev/sdp        100791728    66682228  28989500  70% /home
/dev/sdq        100791728    484632   95187096   1% /var/lib/mysql
/dev/sdr        30237648     692540   28009108   3% /var/lib/pgsql
```

- c. Check the output of `crm_mon` to ensure that the `pm-fs` and `postgresqld` resources are running.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                   Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
ml-fs          (ocf::heartbeat:Filesystem):   Started smw1
cray-syslog    (lsb:cray-syslog):             Started smw1
homedir        (ocf::heartbeat:Filesystem):   Started smw1
md-fs          (ocf::heartbeat:Filesystem):   Started smw1
pm-fs          (ocf::heartbeat:Filesystem):   Started smw1
postgresqld    (lsb:postgresql):             Started smw1
mysqld         (ocf::heartbeat:mysql):        Started smw1
```

Verify the SMW HA Cluster Configuration

Prerequisites

Before beginning this procedure, wait for 30 to 60 seconds for the cluster system to come up after finishing the configuration in the previous procedure.

After rebooting a configured SMW HA system, use this procedure to check that the SMW HA cluster is up and running correctly.

1. Log in as `root` to the active SMW by using the virtual SMW host name (such as `virtual-smw`). After you have logged in successfully, the prompt displays the host name of the active SMW.

NOTE: The examples in this procedure assume that `smw1` is the active SMW.

```
workstation> ssh root@virtual-smw
.
.
.
smw1:~ #
```

2. Verify the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -r1 | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

All resources except `stonith-2` run on the active SMW.

3. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd): Started smw1
fsync (ocf::smw:fsync): Started smw1
hss-daemons   (lsb:rsms): Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
```

```
Resource Group: HSSGroup
ml-fs      (ocf::heartbeat:Filesystem): Started smw1
cray-syslog (lsb:cray-syslog):        Started smw1
homedir    (ocf::heartbeat:Filesystem): Started smw1
md-fs      (ocf::heartbeat:Filesystem): Started smw1
pm-fs      (ocf::heartbeat:Filesystem): Started smw1
postgresqld (lsb:postgresql):                Started smw1
mysqld     (ocf::heartbeat:mysql):    Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

4. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
5. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

Back Up a Newly-installed SMW HA System

After installing and configuring the SMW HA system, create a backup of the SMW, CLE, and SMW HA software.

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is phy7 and is slot 0, and the bootable backup disk is phy6 and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 186; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
/dev/sda
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Physical slot 1:
/dev/sdc
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Physical slot 2:
/dev/sdd
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0
Physical slot 3:
/dev/sdb
/dev/disk/by-id/ata-ST9500620NS_9XF0665V
/dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0
Physical slot 4:
NOT INSTALLED
Physical slot 5:
NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and grub) drive names, the `device.map` mapping file used by grub should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 183 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 183, the `fstab` lines would change from:

```
/dev/sda1 swap                swap                defaults            0 0
/dev/sda2 /                        ext3                acl,user_xattr      1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 swap swap defaults            0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 /      ext3  acl,user_xattr  1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the `grub` bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the `grub` utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the `grub` bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the `grub` utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB  version 0.97  (640K lower / 3072K upper memory)
 [ Minimal BASH-like line editing is supported.  For the first word, TAB
   lists possible command completions.  Anywhere else TAB lists the possible
   completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
  Running "install /boot/grub/stage1 (hd0)  (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
```

```

    Done.
grub> quit

```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
    Boot  Start      End    Blocks  Id System
           63  16771859   8385898+  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
    Boot  Start      End    Blocks  Id System
    * 16771860 312576704 147902422+  83  Linux

```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```

smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSs
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
    Boot  Start      End    Blocks  Id System
           63  16771859   83828   82  Linux
swap / Solaris
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
    Boot  Start      End    Blocks  Id System
    167719 312581807 156207044+  83  Linux

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

```

```

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the phy7 sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors

```

4. Initialize the swap device.

```

smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
(DOS partition table detected). Use -f to force.
Setting up swap space version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed

```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```

smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group

```

```
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872
```

```
Writing inode tables:   done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done
```

This filesystem will be automatically checked every 37 mounts or 180 days, whichever comes first. Use `tune2fs -c` or `-i` to override.

6. Mount the new backup root file system on `/mnt`.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem      1K-blocks      Used Available Use% Mounted on
/dev/sda2        303528624    6438700 281671544   3% /
udev             1030332         116   1030216    1% /dev
/dev/sdb2        306128812    195568 290505224    1% /mnt
```

The running root file system device is the one mounted on `/`.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.
For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
lists possible command completions. Anywhere else TAB lists the
possible
completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
```

```

grub> root (hd2,1)
root (hd2,1)
  Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
  embedded.
  succeeded
  Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1)/boot/grub/
  stage2 \
  /boot/grub/menu.lst"... succeeded
Done.
grub> quit

```

IMPORTANT: For R815 SMWs, grub recreates device.map with the short names, not the persistent names. Do not trust the /dev/sdx names. Always use find when executing grub because it is possible that grub root may not be hd2 the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

(Optional) R815 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device

This optional procedure modifies a bootable backup drive for a Dell R815 SMW in order to boot from and run the R815 SMW from the backup root partition.

IMPORTANT: In order to boot from this backup drive, the primary boot drive must still be operable and able to boot the grub boot blocks installed. If the backup drive is modified to boot as an alternate boot device, it will no longer function as a bootable backup if the primary drive fails.

The disk device names shown in this procedure are only examples. Substitute the actual disk device names for this system. The boot disk is phy7 and is slot 0, and the bootable backup disk is phy6 and is slot 1.

1. Mount the backup drive's root partition.

```
smw# mount /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2 /mnt
```

2. Create a new boot entry in the /boot/grub/menu.lst file. This entry should be a duplicate of the primary boot entry with the following changes:
 - a. Modify the title to uniquely identify the backup boot entry.
 - b. Modify the root (hd0,1) directive to reflect the grub name of the backup drive.
 - c. Modify the root= and resume= specifications to reference the backup drive device.

This is an example /boot/grub/menu.lst file. Note the new entry for the backup drive. This example references phy7 (slot 0) and as the primary drive and phy6 (slot 1) as the backup drive.

```

smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst.20110317
smw# vi /boot/grub/menu.lst
smw# cat /boot/grub/menu.lst
# Modified by YaST2. Last modification on Wed Jun 27 12:32:43 CDT 2012
default 0
timeout 8
##YaST - generic_mbr
gfxmenu (hd0,1)/boot/message
##YaST - activate

###Don't change this comment - YaST2 identifier: Original name: linux###
title SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 \
    resume=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 \
    splash=silent crashkernel=256M-:128M showopts vga=0x31a
    initrd /boot/initrd-3.0.26-0.7-default

### New entry allowing a boot of the back-up drive when the primary drive
### is still present.
title BACK-UP DRIVE - SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd1,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2 \
    resume=/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part1 \
    splash=silent crashkernel=256M-:128M showopts vga=0x31a
    initrd (hd0,1)/boot/initrd-3.0.26-0.7-default

###Don't change this comment - YaST2 identifier: Original name: failsafe###
title Failsafe -- SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 \
    showopts ide=nodma apm=off noresume edd=off powersaved=off \
    nohz=off highres=off processor.max_cstate=1 nomodeset x11failsafe vga=0x31a
    initrd /boot/initrd-3.0.26-0.7-default

```

3. Modify the backup drive's `/etc/fstab` file to reference the secondary drive slot rather than the first drive slot. Examine the backup drive's `fstab` file. Edit the `/mnt/etc/fstab` file, changing `phy7` to `phy6` device names to reference the backup drive. In the following example, the backup drive is `phy6-....`

```

smw# cp -p /mnt/etc/fstab /mnt/etc/fstab.20110317
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
swap swap defaults 0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
/ ext3 acl,user_xattr 1 1
proc /proc proc defaults 0 0
sysfs /sys sysfs noauto 0 0
debugfs /sys/kernel/debug debugfs noauto 0 0
usbfs /proc/bus/usb usbfs noauto 0 0
devpts /dev/pts devpts mode=0620,gid=5 0 0
smw# vi /mnt/etc/fstab

```

```
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
swap swap defaults 0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
/ ext3 acl,user_xattr 1 1
proc /proc proc defaults 0 0
sysfs /sys sysfs noauto 0 0
debugfs /sys/kernel/debug debugfs noauto 0 0
usbfs /proc/bus/usb usbfs noauto 0 0
devpts /dev/pts devpts mode=0620,gid=5 0 0
```

4. Unmount the backup drive.

```
smw# umount /mnt
```

The SMW can now be shut down and rebooted. Upon display of the `Please select boot device` prompt, select the `BACK-UP DRIVE - SLES 11` entry to boot the backup root partition.

R630 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R630 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `pci-0000:03:00.0-scsi-0:0:0:0` and is slot 0, and the bootable backup disk is `pci-0000:03:00.0-scsi-0:0:1:0` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 195; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R630 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```

smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
  /dev/sda
  /dev/disk/by-id/scsi-35000c50079ab34b7
  /dev/disk/by-id/wwn-0x5000c50079ab34b7
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Physical slot 1:
  /dev/sdb
  /dev/disk/by-id/scsi-35000c50079ab71c4
  /dev/disk/by-id/wwn-0x5000c50079ab71c4
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
Physical slot 2:
  /dev/sdc
  /dev/disk/by-id/scsi-35000c50079ab313b
  /dev/disk/by-id/wwn-0x5000c50079ab313b
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
Physical slot 3:
  /dev/sdd
  /dev/disk/by-id/scsi-35000c50079ab4b4c
  /dev/disk/by-id/wwn-0x5000c50079ab4b4c
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
Physical slot 4:
  /dev/sde
  /dev/disk/by-id/scsi-35000c50079d05e70
  /dev/disk/by-id/wwn-0x5000c50079d05e70
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
Physical slot 5:
  NOT INSTALLED
Physical slot 6:
  NOT INSTALLED
Physical slot 7:
  NOT INSTALLED

```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```

smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD

```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and `grub`) drive names, the `device.map` mapping file used by `grub` should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the `grub device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name.

NOTE: `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r630 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
(hd1) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
(hd2) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
(hd3) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
(hd4) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
(hd5) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:5:0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 192 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 183, the `fstab` lines would change from:

```
/dev/sda1 swap                swap                defaults            0 0
/dev/sda2 /                          ext3                acl,user_xattr     1 1
```

to:

```
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1 swap swap
defaults 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 /      ext3
acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the `grub` bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the grub utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the grub utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
 [ Minimal BASH-like line editing is supported. For the first word, TAB
   lists possible command completions. Anywhere else TAB lists the possible
   completions of a device/filename. ]
grub> root (hd0,1)
root (hd0,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
Done.
grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x000c3cc8
```

Blocks	Id	System	Device	Boot	Start	End
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1					2048	67102719
33550336	82	Linux swap / Solaris				
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2	*				67102720	976773119
454835200	83	Linux				

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Command (m for help): d
Partition number (1-4): 2
```

```

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Partition number (1-4, default 1): 1
First sector (2048-976773167, default 2048): (Press the Enter key)
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-976773167, default 976773167): 67102719

Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4, default 2): 2
First sector (67102720-976773167, default 67102720): (Press the Enter key)
Using default value 67102720
Last sector, +sectors or +size{K,M,G} (67102720-976773167, default 976773167): (Press
the Enter key)
Using default value 976773167

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the `pci-0000:03:00.0-scsi-0:0:0:0` sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x7c334e96

          Device Boot      Start         End
Blocks  Id System
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1
33550336  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2
454835224  83  Linux

```

4. Initialize the `swap` device.

```

smw# mkswap /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1
Setting up swapspace version 1, size = 33550332 KiB
no label, UUID=8391498b-d159-469c-b766-66f00a28ff74

```

5. Create a new file system on the backup drive root partition by executing the `mkfs` command.

```
smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
28434432 inodes, 113708806 blocks
5685440 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
3471 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 33 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

6. Mount the new backup root file system on `/mnt`.

```
smw# mount /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem      1K-blocks    Used Available Use% Mounted on
/dev/sda2       447696736  9180648 437606420   3% /
udev            66029308      744  66028564   1% /dev
tmpfs           66029308    39540  65989768   1% /dev/shm
/dev/sdae       309637120 1107516 292800964   1% /var/opt/cray/disk/1
/dev/sdac       206424760 1963664 193975336   2% /home
/dev/sdad       154818540  474696 146479524   1% /var/lib/mysql
/dev/drbd_r0    961405840  247180 912322076   1% /var/lib/pgsql
/dev/sdb2       447696760  202940 424752060   1% /mnt
```

The running root file system device is the one mounted on `/`.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 9129804 blocks.
DUMP: Volume 1 started with block 1 at: Wed Sep 16 15:43:08 2015
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
./tmp/rstidir1442436041: (inode 27254928) not found on tape
./tmp/rstmode1442436041: (inode 27254931) not found on tape
DUMP: 77.64% done at 23626 kB/s, finished in 0:01
```

```
DUMP: Volume 1 completed at: Wed Sep 16 15:50:09 2015
DUMP: Volume 1 9132800 blocks (8918.75MB)
DUMP: Volume 1 took 0:07:01
DUMP: Volume 1 transfer rate: 21693 kB/s
DUMP: 9132800 blocks (8918.75MB)
DUMP: finished in 421 seconds, throughput 21693 kBytes/sec
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015
DUMP: Date this dump completed: Wed Sep 16 15:50:09 2015
DUMP: Average transfer rate: 21693 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the `root` and `swap` devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

```
smw# vi /mnt/etc/fstab
```

For example, change

```
/dev/disk/by-id/scsi-35000c50079ab34b7-part1 swap      swap
defaults                0 0
/dev/disk/by-id/scsi-35000c50079ab34b7-part2 /          ext3
acl,user_xattr          1 1
```

to:

```
/dev/disk/by-id/scsi-35000c50079ab71c4-part1 swap      swap
defaults                0 0
/dev/disk/by-id/scsi-35000c50079ab71c4-part2 /          ext3
acl,user_xattr          1 1
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/scsi-35000c50079ab34b7-part2
```

with:

```
root=/dev/disk/by-id/scsi-35000c50079ab71c4-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to grub boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_1
```

- b. Invoke the grub boot utility. Within the grub boot utility:

1. Execute the `find` command to locate the drive designation that grub uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the grub boot blocks on that drive. The Linux grub utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the
  possible
  completions of a device/filename. ]
grub> find /THIS_IS_1
find /THIS_IS_1
(hd1,1)
grub> root (hd1,1)
root (hd1,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd1)
setup (hd1)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd1)"... 17 sectors are
embedded.
succeeded
Running "install /boot/grub/stage1 (hd1) (hd1)1+17 p (hd1,1)/boot/grub/
stage2 /boot/grub/menu.lst"... succeeded
Done.
grub> quit
quit
```

IMPORTANT: For R630 SMWs, grub recreates `device.map` with the short names, not the persistent names. Do not trust the `/dev/sdx` names. Always use `find` when executing grub because it is possible that grub `root` may not be `hd2` the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

(Optional) R630 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device

This optional procedure modifies a bootable backup drive for a Dell R630 SMW in order to boot from and run the R630 SMW from the backup root partition.

IMPORTANT: In order to boot from this backup drive, the primary boot drive must still be operable and able to boot the grub boot blocks installed. If the backup drive is modified to boot as an alternate boot device, it will no longer function as a bootable backup if the primary drive fails.

The disk device names shown in this procedure are only examples. Substitute the actual disk device names for this system. The boot disk is `pci-0000:03:00.0-scsi-0:0:0:0` and is slot 0, and the bootable backup disk is `pci-0000:03:00.0-scsi-0:0:1:0` and is slot 1.

1. Mount the backup drive's root partition.

```
smw# mount /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /mnt
```

2. Create a new boot entry in the `/boot/grub/menu.lst` file. This entry should be a duplicate of the primary boot entry with the following changes:
 - a. Modify the title to uniquely identify the backup boot entry.
 - b. Modify the `root (hd0,1)` directive to reflect the grub name of the backup drive.
 - c. Modify the `root=` and `resume=` specifications to reference the backup drive device.

This is an example `/boot/grub/menu.lst` file. Note the new entry for the backup drive. This example references `pci-0000:03:00.0-scsi-0:0:0:0` (slot 0) as the primary drive and `pci-0000:03:00.0-scsi-0:0:1:0` (slot 1) as the backup drive.

```
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst.20150916
smw# vi /boot/grub/menu.lst
smw# cat /boot/grub/menu.lst

# Modified by YaST2. Last modification on Thu Aug 13 19:38:47 CDT 2015
default 0
timeout 8
##YaST - generic mbr
gfxmenu (hd0,1)/boot/message
##YaST - activate

###Don't change this comment - YaST2 identifier: Original name: linux###
title SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part2 pci=bfsort resume=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part1 splash=silent crashkernel=256M-:128M@16M
showopts biosdevname=X vga=0x31a
    initrd /boot/initrd-3.0.101-0.46-default

### New entry allowing a boot of the back-up drive when the primary drive
### is still present
title BACK-UP DRIVE - SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd1,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:1:0-part2 pci=bfsort resume=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:1:0-part1 splash=silent crashkernel=256M-:128M@16M
```

```
showopts biosdevname=X vga=0x31a
initrd /boot/initrd-3.0.101-0.46-default
```

```
###Don't change this comment - YaST2 identifier: Original name: failsafe###
title Failsafe -- SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part2 showopts ide=nodma apm=off noresume edd=off
powersaved=off nohz=off highres=off processor.max_cstate=1 nomodeset
x11failsafe biosdevname=X vga=0x31a
    initrd /boot/initrd-3.0.101-0.46-default
```

3. Modify the backup drive's `/etc/fstab` file to reference the secondary drive slot rather than the first drive slot. Examine the backup drive's `fstab` file. Edit the `/mnt/etc/fstab` file, changing `pci-0000:03:00.0-scsi-0:0:0:0` to `pci-0000:03:00.0-scsi-0:0:1:0` device names to reference the backup drive. In the following example, the backup drive is `pci-0000:03:00.0-scsi-0:0:1:0-....`

```
smw# cp /mnt/etc/fstab /mnt/etc/fstab.20150916
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1swap          swap
defaults                0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 /          ext3
acl,user_xattr          1 1
proc                    /proc          proc           defaults       0 0
sysfs                   /sys           sysfs          noauto         0 0
debugfs                 /sys/kernel/debug debugfs        noauto         0 0
usbfs                   /proc/bus/usb  usbfs          noauto         0 0
devpts                  /dev/pts       devpts         mode=0620,gid=5 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0-part1/var/opt/cray/disk/1 ext3
defaults                1 0
none /var/lib/dhcp/db ramfs defaults 0 0

smw# vi /mnt/etc/fstab
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1swap          swap
defaults                0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /          ext3
acl,user_xattr          1 1
proc                    /proc          proc           defaults       0 0
sysfs                   /sys           sysfs          noauto         0 0
debugfs                 /sys/kernel/debug debugfs        noauto         0 0
usbfs                   /proc/bus/usb  usbfs          noauto         0 0
devpts                  /dev/pts       devpts         mode=0620,gid=5 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0-part1/var/opt/cray/disk/1 ext3
defaults                1 0
none /var/lib/dhcp/db ramfs defaults 0 0
```

4. Unmount the backup drive.

```
smw# umount /mnt
```

The SMW can now be shut down and rebooted. Upon display of the Please select boot device prompt, select the `BACK-UP DRIVE - SLES 11` entry to boot the backup root partition.

Change Default HA Passwords After Installation

During HA configuration, the passwords for `hacluster` and the `stonith` resources are set to the `root` password for the SMWs. If you changed the default `root` and iDRAC passwords after installing the SMW software, you do not need to change the passwords again. Otherwise, use the following procedure to change the passwords.

IMPORTANT: The `hacluster`, `stonith`, and iDRAC passwords must be the same as the SMW `root` password. For more information, see [Default Passwords for an SMW HA System](#).

1. Log on to the active SMW (`smw1`) as `root`, using the virtual SMW host name (such as `virtual-smw`). After you have logged in successfully, the prompt displays the host name of the active SMW.

NOTE: The examples in this procedure assume that `smw1` is the active SMW.

2. To change the SMW `root`, `hacluster`, and `stonith` passwords, execute the following commands on `smw1`. The `hacluster` and `stonith` passwords must be the same as the SMW `root` password.

```
smw1:~# passwd root
smw1:~# passwd hacluster
smw1:~# crm resource param stonith-1 set passwd new-passwd
smw1:~# crm resource param stonith-2 set passwd new-passwd
```

3. Change the SMW `root` and `hacluster` passwords on `smw2`, using the same `root` password as on `smw1`. The `hacluster` password must be the same as the `root` password.

```
smw2:~# passwd root
smw2:~# passwd hacluster
```

4. To change the iDRAC passwords, see [Managing System Software for the Cray Linux Environment \(S-2393\)](#). The iDRAC passwords must be the same as the SMW `root` password.

Move Local Log Data to RAID

Prerequisites

New HA software has been installed and administrator may want to preserve pre-existing log data on the local disk.

When an existing non-HA SMW is converted to an HA system, the HA installer creates a new disk on the RAID. This disk contains the directories and debug dump log. The other log files remain on the local disk. If that history is to be retained, it is recommended to copy the information to the RAID.

1. Log on to the active SMW as `root`.
2. Identify the local log disk name by-path from the file `/etc/fstab`. This example uses `/dev/disk/by-path/XX_YY`.

```
#/dev/disk/by-path/XX_YY /var/opt/cray/disk/1 ...
```

-
3. Mount the local log disk.

```
smw1:~ # mount /dev/disk/by-path/XX_YY /mnt
```

4. Copy the local log disk to the RAID log disk.

```
smw1:~ # cp -a /mnt/* /var/opt/cray/disk/1
```

5. Unmount the local disk.

```
smw1:~ # unmount /mnt
```

Customize a Preinstalled SMW HA System

Cray ships SMW HA systems that are completely installed and configured with Cray-specific host names and IP addresses. To complete the configuration on-site, you must reconfigure the system with site-specific IP addresses and host names.

The following tasks are required to configure a preinstalled SMW HA system:

1. [Prepare to Customize an SMW HA System](#) on page 204
2. [Customize the First SMW](#) on page 205
3. [Change the Cluster Configuration on the First SMW](#) on page 206
4. [Customize the Second SMW](#) on page 208
5. [Finish Customizing a Preinstalled SMW HA System](#) on page 209
6. [Verify Cluster Status After Customization](#) on page 210
7. [Back Up a Customized SMW HA System](#) on page 211
8. [Change Default Passwords After Customization](#) on page 230

See these procedures for optional configuration changes:

- [Customize the SMW HA Cluster](#): Change the email address for failover notification, add site-specific files and directories to the synchronization list, and change the migration threshold for SMW HA cluster resources.
- [Configure PMDB Storage](#) on page 172: Configure mirrored or shared storage for the power management database (PMDB).

Prepare to Customize an SMW HA System

Prerequisites

The SMW HA system requires five unique IP addresses and three host names. Determine the following information, as described in [Site-dependent Configuration Values for an SMW HA System](#):

- Virtual host name for the SMW cluster. Users access the SMW HA cluster using this host name. This guide uses the example host name `virtual-smw-default` for the preconfigured host name and `virtual-smw-new` for the site-specific host name.
- Host names of the two SMWs. This guide uses the example host names `smw1-default` and `smw2-default` for the preconfigured host names, and `smw1-new` and `smw2-new` for the site-specific host names.
- Virtual IP address for the SMW cluster.

- IP addresses of the two SMWs. These IP addresses must be on the same subnet as the IP address for the virtual SMW cluster.
- DRAC IP addresses for both SMWs. These IP addresses are used by the iDRAC on each SMW.
- IP addresses for the default gateway and name server.

Before beginning the site customization, the network administrator or site administrator must assign the IP addresses to the corresponding host names for the SMW HA cluster.

The customization procedures update the IP addresses and host names in the following configuration files:

- `/etc/hosts`
- `/etc/hostname`
- `/etc/csync2/csync2.cfg`
- `/etc/csync2/csync2_cray.cfg`
- `/etc/sysconfig/network/ifcfg-eth0`
- `/etc/sysconfig/network/routes`

NOTE: In this procedure, host names and command prompts are shown as `smw1-default` and `smw2-default` before customization. After customization, the SMW host names are shown as `smw1-new` and `smw2-new`.

This procedure starts the customization process.

1. Ensure that the preinstalled system is backed up, as specified in [Back Up a Newly-installed SMW HA System](#).
2. Log in as `root` on each SMW console. Because this procedure changes host names and IP addresses, you **must** execute this procedure on the SMW consoles rather than logging in remotely.
3. Shut down both SMWs, if they are not already shut off.

```
smw1-default:~ # shutdown -h now
```

```
smw2-default:~ # shutdown -h now
```

4. If necessary, connect the Ethernet cables to the network.

Customize the First SMW

Prerequisites

Before beginning this procedure:

- The Ethernet cables must be connected to the network.
- The first SMW must be shut down.

NOTE: In this procedure, host names and command prompts are shown as `smw1-default` and `smw2-default` before customization. After customization, the SMW host names are shown as `smw1-new` and `smw2-new`.

1. Power on the first SMW (`smw1-default`).
2. Log in as `root` on the SMW console. Because this procedure changes host names and IP addresses, you **must** execute this procedure on the SMW console rather than logging in remotely.
3. Execute `yast2` to open the YaST2 Control Center.

```
smw1-default:~ # yast2
```

4. In the right panel, scroll to the Network Devices section and select Network Settings.
5. In the Network Settings window, select the Overview tab.
6. Change the network card setup for the SMW.
 - a. Select `eth0` Customer Network Ethernet, then click the Edit button.
 - b. Enter the IP address of the SMW in the IP Address box.
 - c. Enter the host name of the SMW in the Hostname box.
 - d. Click the Next button to return to the Network Settings window.
7. Define the name servers for the SMW.
 - a. In the Network Settings window, select the Hostname/DNS tab.
 - b. Enter the host name of the SMW in the Hostname box.
 - c. Enter the IP addresses of the name servers into the Name Server boxes. You can define up to three name servers.
 - d. Change the domain name in the Domain Name box to the actual name for the system.
 - e. Change the domain names in the Domain Search box to the actual names for the system.
8. Change the route settings.
 - a. In the Network Settings window, select the Routing tab.
 - b. Enter the IP address for the router in the Default Gateway box.
9. If necessary, change the time zone.
10. To finish the changes, click the OK button. `yast2` writes the configuration changes.
11. Exit `yast2`.

Change the Cluster Configuration on the First SMW

Prerequisites

Before beginning this procedure, log in as `root` on the first SMW's console. Because this procedure changes host names and IP addresses, you **must** execute this procedure on the SMW consoles rather than logging in remotely.

NOTE: In this procedure, host names and command prompts are shown as `smw1-default` and `smw2-default` before customization. After customization, the SMW host names are shown as `smw1-new` and `smw2-new`.

1. As `root` on the first SMW, change the synchronization file `/etc/csync2/csync2.cfg`.

- a. Edit `/etc/csync2/csync2.cfg`.
- b. Locate the following lines in the `ha_group` section:

```
host smw1-default
host smw2-default
```

- c. Change these lines to the actual host names for the system, as in this example:

```
host smw1-new
host smw2-new
```

- d. Save the changes and exit the editor.

2. Change the synchronization file `/etc/csync2/csync2_cray.cfg`.

- a. Edit `/etc/csync2/csync2_cray.cfg`.
- b. Locate the following lines in the `cray_group` section:

```
host smw1-default
host smw2-default
```

- c. Change these lines to the actual host names for the system, as in this example:

```
host smw1-new
host smw2-new
```

- d. Locate the following lines in the `user_group` section:

```
host smw1-default
host smw2-default
```

- e. Change these lines to the actual host names for the system, as in this example:

```
host smw1-new
host smw2-new
```

- f. Save the changes and exit the editor.

3. Change the CRM cluster configuration file.

- a. Edit the cluster configuration file.

```
smw1-default:~ # crm configure edit
```

The configuration file opens in the vi editor.

- b. Locate the following lines.

```
node smw1-default \  
node smw2-default \  
params ip="virtual-smw-default-ip"  
params hostname="smw1-default" ipaddr="drac-smw1-ip-default" userid="root"  
params hostname="smw2-default" passwd="initial0" ipaddr="drac-smw2-ip-default"  
location stonith-1-loc stonith-1 -inf: smw1-default  
location stonith-2-loc stonith-2 -inf: smw2-default
```

- c. Change the host names and IP addresses in these lines to the actual values for the system.

```
node smw1-new \  
node smw2-new \  
params ip="virtual-smw-new-ip"  
params hostname="smw1-new" ipaddr="drac-smw1-ip-new" userid="root"  
params hostname="smw2-new" passwd="initial0" ipaddr="drac-smw2-ip-new"  
location stonith-1-loc stonith-1 -inf: smw1-new  
location stonith-2-loc stonith-2 -inf: smw2-new
```

- d. Save the changes and exit the editor.
4. Shut down `smw1-default`. Wait for the system to finish shutting down before continuing to the next procedure.

Customize the Second SMW

Prerequisites

Before beginning this procedure:

- The `csync2` and CRM cluster configuration files must be customized on the first SMW, as described in [Customize the First SMW](#) on page 205.
- The second SMW must be shut down.

NOTE: In this procedure, host names and command prompts are shown as `smw1-default` and `smw2-default` before customization. After customization, the SMW host names are shown as `smw1-new` and `smw2-new`.

1. Power on the second SMW (`smw2-default`).
2. Log in as `root` on the SMW console. Because this procedure changes host names and IP addresses, you **must** execute this procedure on the SMW console rather than logging in remotely.
3. Execute `yast2` to open the YaST2 Control Center.

```
smw2-default:~ # yast2
```

4. In the right panel, scroll to the Network Devices section and select Network Settings.
5. In the Network Settings window, select the Overview tab.
6. Change the network card setup for the SMW.
 - a. Select eth0 Customer Network Ethernet, then click the Edit button.
 - b. Enter the IP address of the SMW in the IP Address box.
 - c. Enter the host name of the SMW in the Hostname box.
 - d. Click the Next button to return to the Network Settings window.
7. Define the name servers for the SMW.
 - a. In the Network Settings window, select the Hostname/DNS tab.
 - b. Enter the host name of the SMW in the Hostname box.
 - c. Enter the IP addresses of the name servers into the Name Server boxes. You can define up to three name servers.
 - d. Change the domain name in the Domain Name box to the actual name for the system.
 - e. Change the domain names in the Domain Search box to the actual names for the system.
8. Change the route settings.
 - a. In the Network Settings window, select the Routing tab.
 - b. Enter the IP address for the router in the Default Gateway box.
9. If necessary, change the time zone.
10. To finish the changes, click the OK button. `yast2` writes the configuration changes.
11. Exit `yast2`.
12. Shut down `smw2-default`. Wait for the system to finish shutting down before continuing to the next procedure.

Finish Customizing a Preinstalled SMW HA System

Update the cluster configuration to finish customizing the SMW HA system.

NOTE: In this procedure, host names and command prompts are shown as `smw1-default` and `smw2-default` before customization. After customization, the SMW host names are shown as `smw1-new` and `smw2-new`.

1. Power on the first SMW and wait for it to come up. After the system powers on, the prompt displays the new host name (for example, `smw1-new`).

- On the first SMW, remove the default host names from the CRM configuration.

NOTE: In the following commands, replace *smw1-default* with the default (pre-configured) host name of the first SMW. Replace *smw2-default* with the default host name of the second SMW.

```
smw1-new:~ # crm node delete smw1-default
INFO: node smw1-default not found by crm_node
INFO: node smw1-default deleted          =====> deleted

smw1-new:~ # crm node delete smw2-default
INFO: node smw2-default not found by crm_node
INFO: node smw2-default deleted          =====> deleted
```

For each command, ignore the first message that the node is not found. The second message confirms that the node has been deleted.

- Restart the OpenAIS service on the first SMW.

```
smw1-new:~ # rcopenais stop
smw1-new:~ # rcopenais start
```

- Power on the second SMW and wait for it to come up. After the system powers on, the prompt displays the new host name (for example, *smw2-new*).
- Copy the synchronization files */etc/csync2/csync2.cfg* and */etc/csync2/csync2_cray.cfg* from the first SMW to the second SMW.

NOTE: Replace *smw2-new* with the actual host name of the second SMW.

```
smw1-new:~ # scp /etc/csync2/csync2.cfg smw2-new:/etc/csync2/
smw1-new:~ # scp /etc/csync2/csync2_cray.cfg smw2-new:/etc/csync2/
```

- Synchronize the *csync* files between the first SMW and the second SMW.

```
smw1-new:~ # csync2 -xv
```

If all files are synchronized successfully, *csync2* will finish with no errors.

Verify Cluster Status After Customization

Ensure that the SMW HA cluster is operating correctly after changing the cluster configuration.

NOTE: In this procedure, host names and command prompts are shown as *smw1-default* and *smw2-default* before customization. After customization, the SMW host names are shown as *smw1-new* and *smw2-new*.

- As *root* on the first SMW, display the cluster status.

```
smw1-new:~ # crm_mon -1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2-new
Stack: classic openais (with plugin)
Current DC: smw1-new - partition with quorum
```

```
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1-new smw2-new ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1-new
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1-new
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1-new
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1-new
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1-new
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1-new
Notification   (ocf::heartbeat:MailTo):       Started smw1-new
dhcpd (lsb:dhcpd):      Started smw1-new
fsync (ocf::smw:fsync):      Started smw1-new
hss-daemons   (lsb:rsms):      Started smw1-new
stonith-1     (stonith:external/ipmi):       Started smw2-new
stonith-2     (stonith:external/ipmi):       Started smw1-new
Resource Group: HSSGroup
ml-fs        (ocf::heartbeat:Filesystem):    Started smw1-new
cray-syslog  (lsb:cray-syslog):             Started smw1-new
homedir     (ocf::heartbeat:Filesystem):    Started smw1-new
md-fs       (ocf::heartbeat:Filesystem):    Started smw1-new
pm-fs       (ocf::heartbeat:Filesystem):    Started smw1-new
postgresqld (lsb:postgresql):             Started smw1-new
mysqld      (ocf::heartbeat:mysql):        Started smw1-new
```

NOTE: crm_mon may display different resource names, group names, or resource order on the system.

2. Verify that all resources have started. If necessary, see [Verify the SMW HA Cluster Configuration](#) for additional steps to examine cluster status and fix problems with stopped resources or failed actions.

Back Up a Customized SMW HA System

Use the following procedures to back up the customized system.

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is phy7 and is slot 0, and the bootable backup disk is phy6 and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 214; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
  /dev/sda
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Physical slot 1:
  /dev/sdc
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RD7
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Physical slot 2:
  /dev/sdd
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0
Physical slot 3:
  /dev/sdb
  /dev/disk/by-id/ata-ST9500620NS_9XF0665V
  /dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0
Physical slot 4:
  NOT INSTALLED
Physical slot 5:
  NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and `grub`) drive names, the `device.map` mapping file used by `grub` should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 212 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the

numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 212, the `fstab` lines would change from:

```
/dev/sda1 swap                swap        defaults    0 0
/dev/sda2 /                    ext3        acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 swap swap defaults    0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 /      ext3  acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the `grub` bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the `grub` utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the `grub` bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the `grub` utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
  Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

          Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
  Boot  Start      End      Blocks  Id System
          63  16771859   8385898+  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
  Boot  Start      End      Blocks  Id System
  * 16771860 312576704 147902422+  83  Linux
```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
1) software that runs at boot time (e.g., old versions of LILO)
2) booting and partitioning software from other OSs
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080

          Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
  Boot  Start      End      Blocks  Id System
          63  16771859   83828   82  Linux
swap / Solaris
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
  Boot  Start      End      Blocks  Id System
 167719 312581807 156207044+  83  Linux

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
```

```

Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the phy7 sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors

```

4. Initialize the swap device.

```

smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
        (DOS partition table detected). Use -f to force.
Setting up swapspace version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed

```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```

smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872

Writing inode tables:   done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 37 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.

```

6. Mount the new backup root file system on /mnt.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem          1K-blocks      Used Available Use% Mounted on
/dev/sda2            303528624    6438700 281671544   3% /
udev                 1030332         116   1030216   1% /dev
/dev/sdb2            306128812    195568 290505224   1% /mnt
```

The running root file system device is the one mounted on /.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
lists possible command completions. Anywhere else TAB lists the
possible
completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
grub> root (hd2,1)
root (hd2,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
```

```

embedded.
succeeded
  Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1) /boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
Done.
grub> quit

```

IMPORTANT: For R815 SMWs, grub recreates device.map with the short names, not the persistent names. Do not trust the /dev/sdx names. Always use find when executing grub because it is possible that grub root may not be hd2 the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

(Optional) R815 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device

This optional procedure modifies a bootable backup drive for a Dell R815 SMW in order to boot from and run the R815 SMW from the backup root partition.

IMPORTANT: In order to boot from this backup drive, the primary boot drive must still be operable and able to boot the grub boot blocks installed. If the backup drive is modified to boot as an alternate boot device, it will no longer function as a bootable backup if the primary drive fails.

The disk device names shown in this procedure are only examples. Substitute the actual disk device names for this system. The boot disk is phy7 and is slot 0, and the bootable backup disk is phy6 and is slot 1.

1. Mount the backup drive's root partition.

```
smw# mount /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-1un-0-
part2 /mnt
```

2. Create a new boot entry in the /boot/grub/menu.lst file. This entry should be a duplicate of the primary boot entry with the following changes:
 - a. Modify the title to uniquely identify the backup boot entry.
 - b. Modify the root (hd0,1) directive to reflect the grub name of the backup drive.
 - c. Modify the root= and resume= specifications to reference the backup drive device.

This is an example /boot/grub/menu.lst file. Note the new entry for the backup drive. This example references phy7 (slot 0) and as the primary drive and phy6 (slot 1) as the backup drive.

```

smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst.20110317
smw# vi /boot/grub/menu.lst
smw# cat /boot/grub/menu.lst
# Modified by YaST2. Last modification on Wed Jun 27 12:32:43 CDT 2012
default 0
timeout 8
##YaST - generic_mbr

```

```

gfxmenu (hd0,1)/boot/message
##YaST - activate

###Don't change this comment - YaST2 identifier: Original name: linux###
title SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 \
    resume=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 \
    splash=silent crashkernel=256M-:128M showopts vga=0x31a
    initrd /boot/initrd-3.0.26-0.7-default

### New entry allowing a boot of the back-up drive when the primary drive
### is still present.
title BACK-UP DRIVE - SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd1,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2 \
    resume=/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part1 \
    splash=silent crashkernel=256M-:128M showopts vga=0x31a
    initrd (hd0,1)/boot/initrd-3.0.26-0.7-default

###Don't change this comment - YaST2 identifier: Original name: failsafe###
title Failsafe -- SUSE Linux Enterprise Server 11 SP3 - 3.0.26-0.7
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.26-0.7-default \
    root=/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 \
    showopts ide=nodma apm=off noresume edd=off powersaved=off \
    nohz=off highres=off processor.max_cstate=1 nomodeset x11failsafe vga=0x31a
    initrd /boot/initrd-3.0.26-0.7-default

```

3. Modify the backup drive's /etc/fstab file to reference the secondary drive slot rather than the first drive slot. Examine the backup drive's fstab file. Edit the /mnt/etc/fstab file, changing phy7 to phy6 device names to reference the backup drive. In the following example, the backup drive is phy6-....

```

smw# cp -p /mnt/etc/fstab /mnt/etc/fstab.20110317
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
swap swap defaults 0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
/ ext3 acl,user_xattr 1 1
proc /proc proc defaults 0 0
sysfs /sys sysfs noauto 0 0
debugfs /sys/kernel/debug debugfs noauto 0 0
usbfs /proc/bus/usb usbfs noauto 0 0
devpts /dev/pts devpts mode=0620,gid=5 0 0
smw# vi /mnt/etc/fstab
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
swap swap defaults 0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
/ ext3 acl,user_xattr 1 1
proc /proc proc defaults 0 0
sysfs /sys sysfs noauto 0 0
debugfs /sys/kernel/debug debugfs noauto 0 0

```

```
usbfs          /proc/bus/usb      usbfs          noauto          0 0
devpts         /dev/pts           devpts         mode=0620,gid=5 0 0
```

4. Unmount the backup drive.

```
smw# umount /mnt
```

The SMW can now be shut down and rebooted. Upon display of the `Please select boot device` prompt, select the `BACK-UP DRIVE - SLES 11` entry to boot the backup root partition.

R630 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R630 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `pci-0000:03:00.0-scsi-0:0:0:0` and is slot 0, and the bootable backup disk is `pci-0000:03:00.0-scsi-0:0:1:0` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 224; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R630 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
  /dev/sda
  /dev/disk/by-id/scsi-35000c50079ab34b7
```

```

        /dev/disk/by-id/wwn-0x5000c50079ab34b7
        /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Physical slot 1:
        /dev/sdb
        /dev/disk/by-id/scsi-35000c50079ab71c4
        /dev/disk/by-id/wwn-0x5000c50079ab71c4
        /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
Physical slot 2:
        /dev/sdc
        /dev/disk/by-id/scsi-35000c50079ab313b
        /dev/disk/by-id/wwn-0x5000c50079ab313b
        /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
Physical slot 3:
        /dev/sdd
        /dev/disk/by-id/scsi-35000c50079ab4b4c
        /dev/disk/by-id/wwn-0x5000c50079ab4b4c
        /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
Physical slot 4:
        /dev/sde
        /dev/disk/by-id/scsi-35000c50079d05e70
        /dev/disk/by-id/wwn-0x5000c50079d05e70
        /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
Physical slot 5:
        NOT INSTALLED
Physical slot 6:
        NOT INSTALLED
Physical slot 7:
        NOT INSTALLED

```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```

smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD

```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the `grub device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and `grub`) drive names, the `device.map` mapping file used by `grub` should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the `grub device.map` file.

2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name.

NOTE: `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r630 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
(hd1) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
(hd2) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
(hd3) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
(hd4) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
(hd5) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:5:0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 221 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 183, the `fstab` lines would change from:

```
/dev/sda1 swap swap defaults 0 0
/dev/sda2 / ext3 acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1 swap swap
defaults 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 / ext3
acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the grub bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the `grub` utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the grub utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
      Checking if "/boot/grub/stage1" exists... yes
      Checking if "/boot/grub/stage2" exists... yes
      Checking if "/boot/grub/e2fs_stage1_5" exists... yes
      Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
      Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
      grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the fdisk command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x000c3cc8

          Device Boot      Start         End
Blocks   Id System
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1                2048     67102719
33550336  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2      *    67102720     976773119
454835200  83  Linux
```

- b. Use the fdisk command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the d command within fdisk; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type m within fdisk.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
```

```

p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Partition number (1-4, default 1): 1
First sector (2048-976773167, default 2048): (Press the Enter key)
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-976773167, default 976773167): 67102719

Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4, default 2): 2
First sector (67102720-976773167, default 67102720): (Press the Enter key)
Using default value 67102720
Last sector, +sectors or +size{K,M,G} (67102720-976773167, default 976773167): (Press
the Enter key)
Using default value 976773167

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the `pci-0000:03:00.0-scsi-0:0:0:0` sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x7c334e96


```

Blocks	Id	System	Device	Boot	Start	End
33550336	82	Linux swap / Solaris	/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1		2048	67102719
454835224	83	Linux	/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2		67102720	976773167

4. Initialize the swap device.

```

smw# mkswap /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1
Setting up swapspace version 1, size = 33550332 KiB
no label, UUID=8391498b-d159-469c-b766-66f00a28ff74

```

5. Create a new file system on the backup drive root partition by executing the `mkfs` command.

```

smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)

```

```

Fragment size=4096 (log=2)
28434432 inodes, 113708806 blocks
5685440 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
3471 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

```

```

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

```

This filesystem will be automatically checked every 33 mounts or 180 days, whichever comes first. Use tune2fs -c or -i to override.

- Mount the new backup root file system on /mnt.

```
smw# mount /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /mnt
```

- Confirm that the backup root file system is mounted.

```

smw# df
Filesystem      1K-blocks    Used Available Use% Mounted on
/dev/sda2       447696736 9180648 437606420   3% /
udev            66029308     744 66028564    1% /dev
tmpfs           66029308    39540 65989768    1% /dev/shm
/dev/sdae       309637120 1107516 292800964    1% /var/opt/cray/disk/1
/dev/sdac       206424760 1963664 193975336    2% /home
/dev/sdad       154818540 474696 146479524    1% /var/lib/mysql
/dev/drbd_r0    961405840 247180 912322076    1% /var/lib/pgsql
/dev/sdb2       447696760 202940 424752060    1% /mnt

```

The running root file system device is the one mounted on /.

- Dump the running root file system to the backup drive.

```

smw# cd /mnt
smw# dump 0f - /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 9129804 blocks.
DUMP: Volume 1 started with block 1 at: Wed Sep 16 15:43:08 2015
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
./tmp/rstidir1442436041: (inode 27254928) not found on tape
./tmp/rstmode1442436041: (inode 27254931) not found on tape
DUMP: 77.64% done at 23626 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Wed Sep 16 15:50:09 2015
DUMP: Volume 1 9132800 blocks (8918.75MB)
DUMP: Volume 1 took 0:07:01
DUMP: Volume 1 transfer rate: 21693 kB/s
DUMP: 9132800 blocks (8918.75MB)
DUMP: finished in 421 seconds, throughput 21693 kBytes/sec
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015

```

```
DUMP: Date this dump completed: Wed Sep 16 15:50:09 2015
DUMP: Average transfer rate: 21693 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the `root` and `swap` devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

```
smw# vi /mnt/etc/fstab
```

For example, change

```
/dev/disk/by-id/scsi-35000c50079ab34b7-part1 swap      swap
defaults                0 0
/dev/disk/by-id/scsi-35000c50079ab34b7-part2 /          ext3
acl,user_xattr          1 1
```

to:

```
/dev/disk/by-id/scsi-35000c50079ab71c4-part1 swap      swap
defaults                0 0
/dev/disk/by-id/scsi-35000c50079ab71c4-part2 /          ext3
acl,user_xattr          1 1
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/scsi-35000c50079ab34b7-part2
```

with:

```
root=/dev/disk/by-id/scsi-35000c50079ab71c4-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_1
```

- b. Invoke the grub boot utility. Within the grub boot utility:
 1. Execute the `find` command to locate the drive designation that grub uses.
 2. Select the drive to which the boot blocks will be installed with the `root` command.
 3. Use the `setup` command to set up and install the grub boot blocks on that drive. The Linux grub utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the
  possible
  completions of a device/filename. ]
grub> find /THIS_IS_1
find /THIS_IS_1
(hd1,1)
grub> root (hd1,1)
root (hd1,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd1)
setup (hd1)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd1)"... 17 sectors are
embedded.
succeeded
Running "install /boot/grub/stage1 (hd1) (hd1)1+17 p (hd1,1)/boot/grub/
stage2 /boot/grub/menu.lst"... succeeded
Done.
grub> quit
quit
```

IMPORTANT: For R630 SMWs, grub recreates `device.map` with the short names, not the persistent names. Do not trust the `/dev/sdx` names. Always use `find` when executing grub because it is possible that grub `root` may not be `hd2` the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

(Optional) R630 SMW: Set Up the Bootable Backup Drive as an Alternate Boot Device

This optional procedure modifies a bootable backup drive for a Dell R630 SMW in order to boot from and run the R630 SMW from the backup root partition.

IMPORTANT: In order to boot from this backup drive, the primary boot drive must still be operable and able to boot the grub boot blocks installed. If the backup drive is modified to boot as an alternate boot device, it will no longer function as a bootable backup if the primary drive fails.

The disk device names shown in this procedure are only examples. Substitute the actual disk device names for this system. The boot disk is `pci-0000:03:00.0-scsi-0:0:0:0` and is slot 0, and the bootable backup disk is `pci-0000:03:00.0-scsi-0:0:1:0` and is slot 1.

1. Mount the backup drive's root partition.

```
smw# mount /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /mnt
```

2. Create a new boot entry in the `/boot/grub/menu.lst` file. This entry should be a duplicate of the primary boot entry with the following changes:
 - a. Modify the title to uniquely identify the backup boot entry.
 - b. Modify the `root (hd0,1)` directive to reflect the `grub` name of the backup drive.
 - c. Modify the `root=` and `resume=` specifications to reference the backup drive device.

This is an example `/boot/grub/menu.lst` file. Note the new entry for the backup drive. This example references `pci-0000:03:00.0-scsi-0:0:0:0` (slot 0) as the primary drive and `pci-0000:03:00.0-scsi-0:0:1:0` (slot 1) as the backup drive.

```
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst.20150916
smw# vi /boot/grub/menu.lst
smw# cat /boot/grub/menu.lst

# Modified by YaST2. Last modification on Thu Aug 13 19:38:47 CDT 2015
default 0
timeout 8
##YaST - generic_mbr
gfxmenu (hd0,1)/boot/message
##YaST - activate

###Don't change this comment - YaST2 identifier: Original name: linux###
title SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part2 pci=bfsort resume=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part1 splash=silent crashkernel=256M-:128M@16M
showopts biosdevname=X vga=0x31a
    initrd /boot/initrd-3.0.101-0.46-default

### New entry allowing a boot of the back-up drive when the primary drive
### is still present
title BACK-UP DRIVE - SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd1,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:1:0-part2 pci=bfsort resume=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:1:0-part1 splash=silent crashkernel=256M-:128M@16M
showopts biosdevname=X vga=0x31a
    initrd /boot/initrd-3.0.101-0.46-default

###Don't change this comment - YaST2 identifier: Original name: failsafe###
title Failsafe -- SUSE Linux Enterprise Server 11 SP3 - 3.0.101-0.46
    root (hd0,1)
    kernel /boot/vmlinuz-3.0.101-0.46-default root=/dev/disk/by-path/
pci-0000:03:00.0-scsi-0:0:0:0-part2 showopts ide=nodma apm=off noresume edd=off
powersaved=off nohz=off highres=off processor.max_cstate=1 nomodeset
```

```
x11failsafe biosdevname=X vga=0x31a
initrd /boot/initrd-3.0.101-0.46-default
```

3. Modify the backup drive's `/etc/fstab` file to reference the secondary drive slot rather than the first drive slot. Examine the backup drive's `fstab` file. Edit the `/mnt/etc/fstab` file, changing `pci-0000:03:00.0-scsi-0:0:0:0` to `pci-0000:03:00.0-scsi-0:0:1:0` device names to reference the backup drive. In the following example, the backup drive is `pci-0000:03:00.0-scsi-0:0:1:0-...`

```
smw# cp /mnt/etc/fstab /mnt/etc/fstab.20150916
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1swap          swap
defaults                0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 /          ext3
acl,user_xattr           1 1
proc                    /proc          proc           defaults       0 0
sysfs                   /sys           sysfs          noauto         0 0
debugfs                 /sys/kernel/debug debugfs        noauto         0 0
usbfs                   /proc/bus/usb  usbfs          noauto         0 0
devpts                  /dev/pts       devpts         mode=0620,gid=5 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0-part1/var/opt/cray/disk/1 ext3
defaults                1 0
none /var/lib/dhcp/db ramfs defaults 0 0

smw# vi /mnt/etc/fstab
smw# cat /mnt/etc/fstab
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1swap          swap
defaults                0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /          ext3
acl,user_xattr           1 1
proc                    /proc          proc           defaults       0 0
sysfs                   /sys           sysfs          noauto         0 0
debugfs                 /sys/kernel/debug debugfs        noauto         0 0
usbfs                   /proc/bus/usb  usbfs          noauto         0 0
devpts                  /dev/pts       devpts         mode=0620,gid=5 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0-part1/var/opt/cray/disk/1 ext3
defaults                1 0
none /var/lib/dhcp/db ramfs defaults 0 0
```

4. Unmount the backup drive.

```
smw# umount /mnt
```

The SMW can now be shut down and rebooted. Upon display of the `Please select boot device` prompt, select the `BACK-UP DRIVE - SLES 11` entry to boot the backup root partition.

Change Default Passwords After Customization

During HA configuration, the passwords for `hacluster` and the `stonith` resources are set to the `root` password for the SMWs. If you changed the default `root` and `iDRAC` passwords after installing the SMW software, you do not need to change the passwords again. Otherwise, use the following procedure to change the passwords.

IMPORTANT: The `hacluster`, `stonith`, and `iDRAC` passwords must be the same as the SMW `root` password. For more information, see [Default Passwords for an SMW HA System](#).

1. Log on to the active SMW (`smw1`) as `root`, using the virtual SMW host name (such as `virtual-smw`). After you have logged in successfully, the prompt displays the host name of the active SMW.

NOTE: The examples in this procedure assume that `smw1` is the active SMW.

2. To change the SMW `root`, `hacluster`, and `stonith` passwords, execute the following commands on `smw1`. The `hacluster` and `stonith` passwords must be the same as the SMW `root` password.

```
smw1:~# passwd root
smw1:~# passwd hacluster
smw1:~# crm resource param stonith-1 set passwd new-passwd
smw1:~# crm resource param stonith-2 set passwd new-passwd
```

3. Change the SMW `root` and `hacluster` passwords on `smw2`, using the same `root` password as on `smw1`. The `hacluster` password must be the same as the `root` password.

```
smw2:~# passwd root
smw2:~# passwd hacluster
```

4. To change the iDRAC passwords, see *Managing System Software for the Cray Linux Environment (S-2393)*. The iDRAC passwords must be the same as the SMW `root` password.

Configure PMDB Storage

Choose one of these options to configure shared storage for the Power Management Database (PMDB).

- [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172. Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.
- [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. Shared storage: A logical disk, configured as a LUN (Logical Unit) or logical volume on the boot RAID. The boot RAID must have sufficient space for `/var/lib/pgsql`.

Cray strongly recommends using either mirrored storage (preferred) or shared storage. An unshared PMDB is split across both SMWs; data collected before an SMW failover will be lost or not easily accessible after failover. For more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9.

Configure Mirrored Storage with DRBD for the PMDB

Prerequisites

IMPORTANT:

If mirrored storage becomes available after the PMDB has been configured for shared storage, use the procedure [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322 instead of this procedure.

Before beginning this procedure:

- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- Plan sufficient time for this procedure. Transferring the Power Management Database (PMDB) to a 1 TB disk requires about 10 hours. The SMW HA cluster should be in maintenance mode until the synchronization operation completes. The Cray system (compute and service nodes) can remain up and can run jobs during this period.
- Check `/etc/fstab` to ensure that there is no entry for `phy3`.
- If upgrading or updating the SMW HA system, ensure that the following RPMs are installed on both SMWs and that the version number is 8.4.4 or higher:

```
drbd-bash-completion-8.4.4-0.22.9
drbd-kmp-default-8.4.4_3.0.101_0.15-0.22.7
drbd-udev-8.4.4-0.22.9
drbd-utils-8.4.4-0.22.9
drbd-pacemaker-8.4.4-0.22.9
drbd-xen-8.4.4-0.22.9
drbd-8.4.4-0.22.9
```

If necessary, install or update any missing RPMs with `"zypper install drbd"`.

Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.

This procedure configures the network for DRBD, configures the DRBD disks, and transfers the PMDB data from local disk to the mirrored DRBD disks.

1. Add `eth5` to the network files.
 - a. Log in as root on the first SMW (`smw1`).

```
workstation> ssh root@smw1
```

- b. On `smw1`, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.2/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- c. In a separate terminal session, log in as root on the other SMW (`smw2`).

```
workstation> ssh root@smw2
```

- d. On `smw2`, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.3/16'
NAME='eth5 SMW HA DRBD Network'
```

```
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

2. Reinitialize the eth5 interface on both SMWs.

```
smw1:~# ifdown eth5; sleep 1; ifup eth5
```

```
smw2:~# ifdown eth5; sleep 1; ifup eth5
```

3. Verify the IP addresses from *smw1*.

```
smw1:~# ping -c3 10.5.1.3
```

4. Configure the firewall to allow eth5 as an internal connection on both SMWs.

- a. Edit the file `/etc/sysconfig/SuSEfirewall2` on both *smw1* and *smw2*.
- b. Locate the line containing the `FW_DEV_INT` variable.
- c. If necessary, add `eth5` to the end of the `FW_DEV_INT` line.

```
FW_DEV_INT="eth1 eth2 eth3 eth4 eth5 lo"
```

- d. Save your changes and exit the editor on both SMWs.

5. Reinitialize the IP tables by executing the `/sbin/SuSEfirewall2` command on both SMWs.

```
smw1:~# /sbin/SuSEfirewall2
```

```
smw2:~# /sbin/SuSEfirewall2
```

6. On the active SMW only, add the new DRDB disk to the SMW HA configuration.

NOTE: The following examples assume that *smw1* is the active SMW.

- a. Verify that the device exists on both SMWs.
For Dell R-630 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

For Dell R815 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

- b. Determine if the dedicated disk for the PMDB must be formatted. In this procedure, this disk is referred to as `PMDISK`.

NOTE: If the `PMDISK` is already correctly formatted, skip to step [6.f](#) on page 235.

This procedure assumes that a disk drive is available for use as a dedicated drive for the PMDB. The drive should be physically located within the rack-mount SMW at slot 4. The drive should be of the

specification 1 TB 7.2K RPM SATA 3Gbps 2.5in HotPlug Hard Drive 342-1998, per the SMW Bill of Materials. On a Dell PowerEdge R815 the device for PMDISK is /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0 On a Dell PowerEdge R630 the device for PMDISK is /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0.

- c. Verify that the PMDISK is inserted into the SMW by entering the correct device name. This example is for a Dell R815.

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
255 heads, 63 sectors/track, 121601 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1

Device Boot          Start          End          Blocks      Id  System
```

- d. Create a new primary partition for the PMDISK, and write it to the partition table. If there are any existing partitions on this disk, manually delete them first.

```
smw:#fdisk \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)p
Partition number (1-4, default 1): 1
First sector (2048-1953525167, default 2048): [press return]
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-1953525167, default 1953525167): [press
return]
Using default value 1953525167
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

- e. Verify that the partition has been created. This should be device /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
81 heads, 63 sectors/track, 382818 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1

Device Boot          Start          End          Blocks      Id  System
```

```
/dev/disk/by-path/. . .-lun-0-part1      2048 1953525167 976761560 83
Linux
```

- f. Navigate to the directory containing the SMWHAconfig command.

```
smw1:~# cd /opt/cray/ha-smw/default/hainst
```

- g. Execute SMWHAconfig to add the DRBD disk. For *disk-device*, specify the disk ID of the disk backing the DRBD disk, using either the by-name or by-path format for the device name. On a rack-mount SMW (either Dell R815 or R630), the DRBD disk is a partition on the disk in slot 4. On a Dell 815 this is `/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1`. On a Dell 630 it is `/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1`

```
smw1:~# ./SMWHAconfig --add_disk=pm-fs --device=/dev/drbd_r0 --directory=/var/lib/pgsql \
--pm_disk_name=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

7. Reboot the active SMW (*smw1*) and wait for it to boot completely.
8. Reboot the other SMW (*smw2*) and wait for it to boot completely.
9. Correct the permissions for the `/var/lib/pgsql` file on the active SMW.

```
smw1:~# chown postgres:postgres /var/lib/pgsql
smw1:~# chmod 750 /var/lib/pgsql
```

10. Put the SMW HA cluster into maintenance mode while waiting for the DRBD sync operation to complete. When *smw1* and *smw2* rejoin the cluster after rebooting, the primary DRBD disk (in *smw1*) synchronizes data to the secondary disk (in *smw2*). DRBD operates at the device level to synchronize the entire contents of the PMDB disk. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. The Cray system (service and compute nodes) can be booted and can run jobs during this period.

IMPORTANT:

Cray strongly recommends putting the SMW HA cluster into maintenance mode to prevent any failover during the sync operation. If a failover were to occur during this period, the newly-active SMW could have an incomplete copy of PMDB data.

- a. Put the SMW HA cluster into maintenance mode on *smw1*.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- b. Check the status of the DRBD sync operation with either `rdrbd status` or `cat /proc/drbd`. The `rdrbd` output is easier to read, but `/proc/drbd` contains more status information and includes an estimate of time to completion.

```
smw1:~# rdrbd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res  cs          ro          ds          p
mounted          fstype
0:r0  SyncSource  Primary/Secondary  UpToDate/Inconsistent  C  /var/lib/
pgsql  ext3
...   sync'ed:    72.7%          (252512/922140)M
```

```

smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
 0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
    ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2
ua:0 ap:0 ep:1 wo:f oos:260636656
    [=====>.....] sync'ed: 72.4% (254524/922140)M
    finish: 2:21:07 speed: 30,768 (29,720) K/sec

```

For an explanation of the status information in `/proc/drbd`, see the DRDB User's Guide at [linbit.com: http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd](http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd).

- When the DRBD sync operation finishes, bring the HA cluster out of maintenance mode on `smw1`.

```

smw1:~# crm configure property maintenance-mode=false 2> /dev/null

```

- Examine the output of `crm status` to ensure that the `ip_drbd_pgsql` is started on `smw1` and that the Masters and Slaves entries for `ms_drbd_pgsql` display the SMW host names (`smw1` and `smw2`).

```

smw1:~# crm status
Last updated: Thu Jan 22 18:40:21 2015
Last change: Thu Jan 22 11:51:36 2015 by hacluster via crmd on smw1
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.11-3ca8c3b
2 Nodes configured, 2 expected votes
23 Resources configured

Online: [ smw1 smw2 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
.
.
.
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysql        (ocf::heartbeat:mysql):        Started smw1
  ip_drbd_pgsql (ocf::heartbeat:IPaddr2):      Started smw1
Master/Slave Set: ms_drbd_pgsql [drbd_pgsql]
Masters: [ smw1 ]
Slaves: [ smw2 ]

```

Configure Shared Storage on the Boot RAID for the PMDB

Prerequisites

The SMW HA system can be configured to store the Power Management Database (PMDB) on shared storage, a logical disk configured as a LUN (Logical Unit) or logical volume on the boot RAID.

IMPORTANT: Cray strongly recommends using mirrored storage, if available, for the PMDB; for more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9. To move the PMDB from shared storage to mirrored storage, see [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322.

Before beginning this procedure:

- Ensure that the boot RAID contains a LUN for the PMDB with sufficient space for the data. Use the following command to check the size of `/var/lib/pgsql` on the local disk:

```
smw1:~ # du -hs /var/lib/pgsql
```

- Check that the boot RAID is connected.
- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- To capture typescript output from this procedure, do not use a typescript session running directly on the SMW. To save the output of this procedure, use the `script` command to start the typescript session on your local workstation before logging into the SMW, as in this example:

```
workstation> script -af my_output_file
Script started, file is my_output_file
workstation> ssh crayadm@smw1
```

Use this procedure to configure the RAID disk and transfer the power management data base (PMDB) to the power management disk on the shared boot RAID.

- Shut down the Cray system by typing the following command as `crayadm` on the active SMW (`smw1`).

```
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
```

- Log into the active SMW as `root`, either at the console or by using the actual (not virtual) host name.

IMPORTANT: You must log in directly as `root`. Do not use `su` from a different SMW account such as `crayadm`.

- Change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

- Use the `SMWHAconfig` command to move the PMDB and configure the required HA resources. In the following command, replace `scsi-xxxxxxxx` with the persistent device name for the PMDB directory on the boot RAID.

```
smw1:~ # ./SMWHAconfig --add_disk=pm-fs \
--device=/dev/disk/by-id/scsi-xxxxxxxx --directory=/var/lib/pgsql
```

This command mounts the PMDB directory (`/var/lib/pgsql`) to the boot RAID, copies the PMDB data, and configures the HA resources `pm-fs` and `postgresltd`.

5. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

6. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

7. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons   (lsb:rsms):                    Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir     (ocf::heartbeat:Filesystem):    Started smw1
  md-fs       (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

8. Verify that the Power Management Database is on the boot RAID and that the required PMDB resources are running.
 - a. Examine the log file `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out` to verify that the Power Management Database disk appears in the `Cluster RAID Disks` section (at the end of the file), as in this example.

```
----- Cluster RAID Disks -----
07-07 20:47 INFO  MySQL Database disk = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  Log disk             = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  /home disk          = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  PM database disk    = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  ***** Ending of HA software add_disk *****
```

- b. Ensure that the power management file system is mounted by checking for `/var/lib/pgsql` in the output of the `df` command.

```
smw1:~ # df
Filesystem      1K-blocks      Used Available Use% Mounted on
/dev/sda2       120811676    82225412  32449332  72% /
udev            16433608         756  16432852   1% /dev
tmpfs           16433608     37560  16396048   1% /dev/shm
/dev/sdo        483807768  197536596  261695172  44% /var/opt/cray/disk/1
/dev/sdp        100791728    66682228  28989500  70% /home
/dev/sdq        100791728     484632   95187096   1% /var/lib/mysql
/dev/sdr        30237648     692540   28009108   3% /var/lib/pgsql
```

- c. Check the output of `crm_mon` to ensure that the `pm-fs` and `postgresld` resources are running.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
```

```
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons   (lsb:rsms):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog (lsb:cray-syslog):             Started smw1
  homedir    (ocf::heartbeat:Filesystem):    Started smw1
  md-fs     (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs     (ocf::heartbeat:Filesystem):    Started smw1
  postgresql (lsb:postgresql):             Started smw1
  mysqld    (ocf::heartbeat:mysql):     Started smw1
```

Upgrade the Cray SMW HA System

Cray provides periodic upgrades to the SMW, CLE, and SMW HA software releases, as well as infrequent upgrades to the base operating system version running on the SMW. A software upgrade on an SMW HA system involves installing the next major release for all software components. In an upgrade release, the major version number changes. For example, if your system is currently running the CLE 5.1, SMW 7.1, and SMW HA SLEHA 11 SP2 release, you can upgrade to the CLE 5.1, SMW 7.2, and SMW SLEHA 11 SP3 base releases.

Note the requirements for upgrading an SMW HA system:

- Upgrade to the SUSE Linux Enterprise Server version 11 Service Pack 3 (SLES 11 SP3) SMW base operating system before upgrading the SMW and SMW HA software.
- Upgrade the base operating system software, SMW software, and SMW HA software in the same upgrade session. Cray recommends upgrading in the following order:
 1. Operating system software
 2. SMW software
 3. SMW HA software
- For each upgrade release package, upgrade the active SMW first, then upgrade the passive SMW. Do not upgrade both SMWs at the same time.
- The CLE software can be upgraded either before or after the OS, SMW, and SMW HA upgrades.

The upgrade procedures use the following conventions to refer to the SMWs:

- The host name `smw1` specifies the currently active SMW. In examples, the prompt `smw1:~ #` shows a command that runs on this SMW.
- The host name `smw2` specifies the currently passive SMW. In examples, the prompt `smw2:~ #` shows a command that runs on this SMW.
- The host name `virtual-smw` host name specifies the active SMW (which could be either `smw1` or `smw2`). This virtual host name was defined during initial installation.

Before You Start an SMW HA Upgrade

To prepare to upgrade the software on an SMW HA system, do the following:

- Read the *SMW HA Release Notes* and the *SMW HA README* to confirm the required versions for the operating system, SMW, and SMW HA software release, as well as the supported upgrade or update paths. These documents are provided with the SMW HA release package.
- Read the *SMW HA README* and *SMW HA Errata* to determine whether there are any corrections to the upgrade or update procedures. These documents are included in the SMW SLEHA upgrade or update directory.
- Read any Field Notices (FNs) related to kernel security fixes.

IMPORTANT: Kernel 3.0.101-0.461 (provided in FN-6029) or later is required for an SMW HA system. There is a SLES kernel dependency on the `ocfs2-kmp-default` RPM package that will prevent some SLES HA RPMs from being installed unless this kernel update has been applied.

- Back up the current SMW and SMW HA software before installing the upgrade or update packages. For more information, see [R815 SMW: Create an SMW Bootable Backup Drive](#) on page 182.
- Cray recommends checking all file systems with `fsck` before beginning an upgrade or update, because an SMW HA system requires several reboots during this procedure. Exceeding a file-system mount count would delay a reboot by triggering an automatic file-system check.
- Identify any local changes to the list of synchronized files and directories in `/etc/csync2/csync2_cray.cfg`. The installation procedure saves local changes in a temporary file. You will restore those changes in a post-installation step.
- Plan sufficient time. An SMW HA system requires more time to upgrade or update, as compared to a system with a single SMW, because you must install the software on both SMWs. Allow at least two hours of additional time.

Upgrade the Operating System Software

For a system running SLEHA 11 SP2, you must upgrade to the SLES 11 SP3 operating system before upgrading the SMW and SMW HA software.

Upgrade the active SMW first, then upgrade the passive SMW. You must complete the upgrade on the active SMW before starting to upgrade the passive SMW.

NOTE: During this procedure, you will need to refer to the operating system upgrade procedures in [Upgrading the SMW Base Operating System to SLES 11 SP3 \(S-0047\)](#).

IMPORTANT: FN-6029 is required for an SMW HA system. There is a SLES kernel dependency on the `ocfs2-kmp-default` RPM package that will prevent some SLES HA RPMs from being installed unless this FN has been applied.

1. Log on to both SMWs as `root`.
2. Find the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -1 | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

NOTE: The examples in this procedure assume that `smw1` is currently the active SMW.

3. Upgrade the active SMW (`smw1`) by following the applicable procedures in [Upgrading the SMW Base Operating System to SLES 11 SP3](#). This step summarizes the tasks required to upgrade the operating system on `smw1`.
 - a. Back up the current software.
 - b. Shut down the Cray system.

IMPORTANT: Ensure that the boot RAID is powered off or disconnected before continuing.

- c. Upgrade the SMW base operating system to SLES 11 SP3. Follow the procedure for a rack-mount SMW.

The installation process automatically reboots the SMW to finish setting up SP3. After the reboot completes, reconnect the boot RAID to the SMW, then reboot the SMW again to ensure that the boot RAID connection is recognized correctly.

IMPORTANT: You must complete the operating system upgrade on `smw1` before you start the upgrade on `smw2`.

TIP: After the operating system has been upgraded on `smw1`, Cray recommends upgrading the operating system on `smw2` before upgrading the SMW, CLE, and SMW HA software. If necessary, however, you can complete all SMW, CLE and SMW HA software installation on `smw1` before installing the software on `smw2`. Wait until all software is installed on both SMWs before configuring the upgraded SMW HA cluster. Do not upgrade the SMW HA cluster configuration (as described in [Upgrade SMW HA Software](#) on page 285) until the SMW HA update release has been installed on both SMWs.

4. Upgrade the other SMW (`smw2`) by following the applicable procedures in Upgrading the SMW Base Operating System to SLES 11 SP3. This step summarizes the tasks required to upgrade the operating system on `smw2`.
 - a. Back up the current software.
 - b. Skip the step to shut down the Cray system. This step was already done on `smw1`.
 - c. Upgrade the operating system to SLES 11 SP3. Follow the procedure for a rack-mount SMW.

The installation process automatically reboots the SMW to finish setting up SP3. After the reboot completes, reconnect the boot RAID to the SMW, then reboot the SMW again to ensure that the boot RAID connection is recognized correctly.

5. When you have confirmed that the upgrade was successful, create a single bootable backup drive as described in appendix A of Upgrading the SMW Base Operating System to SLES 11 SP3.

Upgrade the SMW Software

Before upgrading the SMW software, ensure that the operating system has been upgraded. For a system running the SMW 7.1 release software, the first upgrade to the SMW 7.2 release package requires upgrading the operating system before upgrading the SMW release software.

The following procedures are required:

1. [Prepare the SMW HA System for an SMW Upgrade](#) on page 244
2. [Upgrade SMW Software on the Active SMW](#) on page 245 (including steps to correct the MySQL path)
3. [Upgrade SMW Software on the Passive SMW](#) on page 262
4. [Finish the SMW Upgrade](#) on page 282

When upgrading the SMW software, upgrade the active SMW first, then upgrade the passive SMW. Do not upgrade both SMWs at the same time.

NOTE:

If the system is already running the SMW 7.2.UP00 (or later) release software, use the procedures in [Update the Cray SMW HA System](#).

Prepare the SMW HA System for an SMW Upgrade

Prerequisites

Before updating the SMW software, ensure that the operating system has been updated to the required release. For more information, see [Upgrade the Operating System Software](#) on page 242

1. Log on to each SMW as `root`.
2. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, shut down the VNC server.

```
smw1:~ # /etc/init.d/vnc stop
```

3. Determine whether the `postgresql` service is currently on or off, and record this state. After completing the upgrade or update, you will return the `postgresql` service to the same state.

```
smw1:~ # chkconfig postgresql
postgresql state
```

4. Find the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -1 | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

NOTE: The examples in this procedure assume that `smw1` is currently the active SMW.

5. Record the iDRAC IP address of both SMWs in case you need to power-cycle either SMW.

Usually, the iDRAC host name follows the naming convention `hostname-drac`. For example, if the host names are `smw1` and `smw2`, the iDRAC host names would be `smw1-drac` and `smw2-drac`. Use the following ping commands to display the iDRAC IP addresses.

NOTE: In these commands, replace `smw1-drac` with the host name of the iDRAC on the active SMW. Replace `smw2-drac` with the host name of the iDRAC on the passive SMW.

```
smw1:~ # ping smw1-drac
PING smw1-drac.us.cray.com (172.31.73.77) 56(84) bytes of data.
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=1 ttl=64
time=1.85 ms
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=2 ttl=64
time=0.398 ms
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=3 ttl=64
time=0.408 ms
...
```

```
smw1:~ # ping smw2-drac
PING smw2-drac.us.cray.com (172.31.73.79) 56(84) bytes of data.
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=1 ttl=64
time=1.85 ms
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=2 ttl=64
time=0.398 ms
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=3 ttl=64
```

```
time=0.408 ms
...
```

- Shut down the Cray system as `crayadm` on the active SMW.

```
smw1:~ # su - crayadm smw1
...
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
...
crayadm@smw1:~> exit
smw1:~ #
```

- As `root` on the active SMW, stop file synchronizing.

```
smw1:~ # crm resource stop fsync
```

- On the active SMW, turn on maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=true 2> /dev/null
```

- Determine the persistent device name of the `/home` directory.

```
smw1:~ # crm configure show | grep device | awk '{print $2 " " $3}' | \
sed 's/"//g'

device=/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9c01 directory=/home
device=/dev/disk/by-id/scsi-360080e500023bff6000006b3515d9bdf directory=/var/
lib/mysql
device=/dev/disk/by-id/scsi-360080e500023bff6000006b1515d9bc9 directory=/var/
opt/cray/disk/1
```

Upgrade SMW Software on the Active SMW

Use the following procedures to to upgrade the SMW software on the active SMW.

Update or Upgrade the Cray SMW Software

Cray provides periodic updates to each System Management Workstation (SMW) release, as well as upgrade releases. In an upgrade release, the major and/or minor version number will change, for example from 7.1.UP01 to 7.2.UP00. In an update release, only the minor version (numbers following *UP*) will change.

Follow the procedures in this chapter to install an SMW 7.2UP04 package. The procedures provided in this chapter do not change the base operating system version running on the SMW.



CAUTION: The SMW must be running the SUSE Linux Enterprise Server version 11 Service Pack 3 (SLES 11 SP3) SMW base operating system and a release of SMW 7.1 or later in order to perform the following update/upgrade procedures.

Prepare to Upgrade or Update SMW Software

IMPORTANT: For a system configured for SMW high availability (HA) with the SMW failover feature, prepare both SMWs for an upgrade.

- Determine which SLES version is running on the SMW by executing the following command:

```
crayadm@smw> cat /etc/SuSE-release
```

- Read the *SMW README* and *SMW Errata* provided in the SMW update directory for any changes to upgrade or update procedures.
- Read the Field Notices (FN) related to kernel security fixes, and apply any needed changes before continuing with the installation.
- If local changes have been made to the file `/opt/cray/hss/default/etc/sedc_srv.ini`, note the following information.
 - Cray software manages this file as a symbolic link to `/opt/cray/hss/default/etc/sedc_srv.ini.xtek` (Cray XE and Cray XK systems only) or to `/opt/cray/hss/default/etc/sedc_srv.ini.cascade` (Cray XC30 systems only). The following actions are taken during software updates:
 - If the symbolic link exists, it is not altered.
 - If the symbolic link does not exist, it is created as specified above.
 - If `sedc_srv.ini` exists not as a symbolic link but as a regular file, it is renamed to `/opt/cray/hss/default/etc/sedc_srv.ini-YYYYMMDDhhmmss` (where `YYYYMMDDhhmmss` is the date and time the file was renamed) and a new symbolic link is created.
 - Before beginning the upgrade, copy `sedc_srv.ini` to a new site-specific file and change the symbolic link to point to that file. After the software update, compare the local file to the distribution `sedc_srv.ini.xtek` or `sedc_srv.ini.cascade` file for any changes that should be merged into the locally modified file.
- If local changes have been made to any automation files, such as `/opt/cray/hss/default/etc/auto.xtshutdown`, back them up before beginning the SMW upgrade.
- For Cray XC Systems: To retain any power management profiles that were created, back up all of the files in the `/opt/cray/hss/7.1.0/pm/profiles` directory before beginning the SMW upgrade.
- For liquid-cooled Cray XC30 Systems only: One or more patches to the SMW 7.1.UP01 release may have installed an `hss.ini` file on the system. This file, and a file containing the default values, `hss.ini.dist` are located in `/opt/tftpboot/ccrd`.
 - If `hss.ini` does not have local changes, delete this file before beginning the update installation, and an active `hss.ini` file will be created as part of the installation.
 - If `hss.ini` has local changes, save a local copy of `hss.ini` to another location. After installation is complete, copy `hss.ini.dist` to `hss.ini` and re-create the local changes in the new `hss.ini` file.
- If using the Cray simple event correlator (SEC) and the `/opt/cray/default/SEC_VARIABLES` file has local changes, make a backup copy of this file before beginning the upgrade or update. For more information, see *Configure Cray SEC Software (S-2542)*.
- If `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` has local changes, the local changes are saved during the upgrade procedure. Also, if `/opt/cray/hss/default/etc/xtnlrd.ini` has local changes and the new release includes an updated `xtnlrd.ini` file, the local version of the file is preserved during the upgrade and the new file is installed as `xtnlrd.ini.rpmnew`. After the upgrade, compare the two files and merge any changes into the local `xtnlrd.ini` file.
- Update the `properties.local` file. (See [Update the properties.local File](#))
- Back up the current software. (See [Back Up the Current Software](#) on page 24).

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is phy7 and is slot 0, and the bootable backup disk is phy6 and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 250; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
/dev/sda
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Physical slot 1:
/dev/sdc
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Physical slot 2:
/dev/sdd
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0
Physical slot 3:
/dev/sdb
/dev/disk/by-id/ata-ST9500620NS_9XF0665V
/dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0
Physical slot 4:
NOT INSTALLED
Physical slot 5:
NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and grub) drive names, the `device.map` mapping file used by grub should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 247 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 247, the `fstab` lines would change from:

```
/dev/sda1 swap                swap        defaults    0 0
/dev/sda2 /                      ext3        acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 swap swap defaults    0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 /      ext3  acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the `grub` bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the `grub` utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the `grub` bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the `grub` utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB  version 0.97  (640K lower / 3072K upper memory)
 [ Minimal BASH-like line editing is supported.  For the first word, TAB
   lists possible command completions.  Anywhere else TAB lists the possible
   completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
  Running "install /boot/grub/stage1 (hd0)  (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
```

```

    Done.
grub> quit

```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
    Boot  Start      End    Blocks  Id System
           63  16771859   8385898+  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
    Boot  Start      End    Blocks  Id System
    * 16771860 312576704 147902422+  83  Linux

```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```

smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSs
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
    Boot  Start      End    Blocks  Id System
           63  16771859   83828   82  Linux
swap / Solaris
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
    Boot  Start      End    Blocks  Id System
    167719 312581807 156207044+  83  Linux

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

```

```

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the phy7 sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors

```

4. Initialize the swap device.

```

smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
(DOS partition table detected). Use -f to force.
Setting up swapspace version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed

```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```

smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group

```

```
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872
```

```
Writing inode tables:   done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done
```

```
This filesystem will be automatically checked every 37 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

6. Mount the new backup root file system on /mnt.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem            1K-blocks      Used Available Use% Mounted on
/dev/sda2              303528624    6438700 281671544   3% /
udev                  1030332         116   1030216   1% /dev
/dev/sdb2              306128812    195568 290505224   1% /mnt
```

The running root file system device is the one mounted on /.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's fstab and menu.lst files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.
For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
lists possible command completions. Anywhere else TAB lists the
possible
completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
```

```

grub> root (hd2,1)
root (hd2,1)
  Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
  embedded.
  succeeded
  Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1)/boot/grub/
  stage2 \
  /boot/grub/menu.lst"... succeeded
Done.
grub> quit

```

IMPORTANT: For R815 SMWs, grub recreates device.map with the short names, not the persistent names. Do not trust the /dev/sdx names. Always use find when executing grub because it is possible that grub root may not be hd2 the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

Shut Down the Cray System

1. Log on to the SMW as `crayadm` and confirm the Cray system is shut down.

```
crayadm@smw:~> ping boot
```

If the command responds with "alive", then it is up and needs to be shut down.

2. Shut down the system by typing the following command.

```
crayadm@smw> xtbootsys -s last -a auto.xtshutdown
```

For more information about using automation files, see the `xtbootsys(8)` man page.

Update the SMW Software and Configuration

1. Open a terminal window, and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. Mount the release media by using one of the following commands, depending on the media type.
 - To install the update package from disk, place the SMW 7.2UP04 Software DVD in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as `/tmp/SMW_version` on the SMW and then substitute this path for `/media/cdrom` in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the `smw-image` ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

3. If `postfix` is not configured on the SMW, skip this step:

To prevent the `master.cf` and `main.cf` `postfix` configuration files from being recreated during software updates or fixes, ensure the following setting in the `/etc/sysconfig/mail` file on the SMW is set to "no":

```
MAIL_CREATE_CONFIG="no"
```

4. To see what SMW software will be updated in this new release, execute these commands prior to doing the update. This information is gathered, displayed, and contained in the log files during the `SMWinstall` process.

- a. Check security and recommended updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GS
```

- b. Check Cray software updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GV
```

5. Create a new copy of the `SMWinstall.conf` configuration file and modify the new copy of the `SMWinstall.conf` file with site-specific requirements. Only `root` can modify the `SMWinstall.conf` configuration file. The `SMWinstall.conf` configuration file is created during the installation process by copying the `SMWinstall.conf` template from the distribution media. By default, the SMW configuration file is placed in `/home/crayadm/SMWinstall.conf`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
smw# chmod 644 /home/crayadm/SMWinstall.conf
smw# vi /home/crayadm/SMWinstall.conf
```

IMPORTANT: For an SMW HA system, define this variable as if the system was a standalone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.

For a description of the contents of the `SMWinstall.conf` file, see the `SMWinstall.conf(5)` man page.

6. Update the software with `SMWinstall`. `SMWinstall` checks for any inconsistency between the system and the `SMWinstall.conf` file settings, prompts for the root MySQL database password, and stores its log files in `/var/adm/cray/logs`.

```
smw# /media/cdrom/SMWinstall
...
```

```
Please enter your root DB password:
Please confirm your root DB password:
Password confirmed.
```

When `SMWinstall` finishes, it will suggest a reboot of the SMW.

7. If necessary, restore the locally modified versions of the following files.

- a. If the installation site had locally modified versions of `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` before this SMW update, restore the local modifications to these files. During the upgrade procedure, the old files are saved in `/etc/syslog-ng/syslog-ng.conf-YYYYMMDDhhmm` and `/etc/rsyslog.conf.rpmsave`.
 - b. If the installation site had local modifications to `/opt/cray/hss/default/etc/auto.xtshutdown` before this SMW update, restore the local modifications to this file.
 - c. If the installation site had power management profiles, copy the files that were backed up from the `/opt/cray/hss/7.1.0/pm/profiles` directory into the new `/opt/cray/hss/default/pm/profiles` directory. See *Monitoring and Managing Power Consumption on the Cray XC30 System (S-0043)* for instructions on validating the restored profiles.
 - d. If the installation site had local modifications to `/opt/cray/hss/default/etc/sedc_srv.ini` before this SMW update, locate the destination of this symbolic link (see [Prepare to Upgrade or Update SMW Software](#) on page 245), compare the content of the local file to the distributed version of the file, and update the local file appropriately.
8. If the installation site has an `/opt/cray/hss/default/etc/xtdiscover.ini` file, the SMW update process does not overwrite an existing `xtdiscover.ini` file; the new version is named `xtdiscover.ini.dist`. Compare the content of the new `xtdiscover.ini.dist` with the original `xtdiscover.ini` file, and update the `xtdiscover.ini` file appropriately.

NOTE: If the `xtdiscover.ini` file does not exist, then the `xtdiscover.ini.dist` file is copied to the `xtdiscover.ini` file.

9. Unmount the SMW 7.2UP04 Software media.

```
smw# umount /media/cdrom
```

10. If using the update disk, eject the SMW 7.2UP04 Software DVD.

```
smw# eject
```

11. Reboot the SMW.

```
smw# reboot
```

For a Cray XE or Cray XK system, continue to [Update the L0 and L1 Firmware](#).

Correct the MySQL Path and Change SMW Status

Prerequisites

After rebooting the active SMW in the previous procedure, wait until the SMW has finished rebooting and all cluster services have started.

Before continuing with the SMW software upgrade, correct the MySQL path, put the passive SMW into standby mode, and take the active SMW out of maintenance mode.

NOTE:

The examples in this procedure assume that `smw1` was the active SMW at the start of the upgrade.

1. Log into each SMW as `root`.

2. Change the SMW HA `mysqld` resource to use the new path for MySQL 5.5.

- a. Display the current paths for the `mysqld` resource.

```
smw1:~ # crm resource param mysqld show binary
/opt/MySQL/default/sbin/mysqld
smw1:~ # crm resource param mysqld show client_binary
/opt/MySQL/default/bin/mysql
```

- b. Change to the new paths for the server and client.

```
smw1:~ # crm resource param mysqld set binary /usr/sbin/mysqld
smw1:~ # crm resource param mysqld set client_binary /usr/bin/mysql
```

- c. Verify the changes.

```
smw1:~ # crm resource param mysqld show binary
/usr/sbin/mysqld
smw1:~ # crm resource param mysqld show client_binary
/usr/bin/mysql
```

3. Put the passive SMW into standby mode.

NOTE: Replace `smw2` with the host name of the passive SMW.

```
smw2:~ # crm node standby smw2
```

4. On the active SMW (`smw1`), turn off maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=false 2> /dev/null
```

For Cray XC Series Systems Only: Update the BC and CC Firmware

Cray XC Series images for the cabinet controller (CC) and blade controller (BC) are always downloaded over the HSS network. Any updated firmware will be placed in `/opt/cray/hss-images/...` as part of the installation or update process. In order to boot the updated firmware, the `hss_make_default_initrd` script must be run, all CCs rebooted, and all BCs rebooted and power cycled.

IMPORTANT: If an installation step fails because of a hardware issue, such as a cabinet failing to power up, when that issue is resolved, go back to the last successful step in the installation procedure and continue from there. Do not skip steps or continue out of order.

1. Update the controller boot image.

The version used in the command argument for `hss_make_default_initrd` should match that of the version specified in the `lsb-cray-hss` line in the output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Mon Sep 14 00:57:20 CDT 2015 on hssbld0 by bwdev
lsb-cray-hss-7.2.0-1.0702.37336.662
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.37336.662
::: Verifying base RPM list to the manifest
::: Installing filesystem hierarchy
::: Installing busybox
::: Installing base RPMs
::: Installing ssh
```

```

::: Removing /etc/securetty installed by the pam RPM
::: Installing sshfs
::: Installing rsh
::: Modifying /etc/pam.d/rlogin to remove securetty checking
::: Modifying /etc/pam.d/rsh to remove rhosts and nologin checking (Bug #779466)
::: Installing rsync
::: Installing atftp
::: Installing tcpdump
::: Installing ethtool
::: Installing syslog-ng
::: Installing logrotate
::: Installing ntp/ntpd
::: Installing strace
::: Installing screen
::: Installing minicom
::: Installing ppp
::: Installing mtd-utils
::: Installing /init
::: Installing file.rpm
::: Installing libgmodule
::: Installing Midnight Commander
::: Installing cray-viper
::: Installing spread
::: Installing coreboot-utils
::: Clearing init.d to be replaced by cray-hss32-filesystem
::: Creating initial etc files needed for root creation
::: Installing Cray kernel
::: Installing latest Cray kernel modules
::: Clearing select /boot items
::: Installing boot-parameters
::: Installing cray-hss32-scripts
::: Installing lsb-cray-hss-controllers
::: Installing cray-libconfig
::: Installing cray-bdm
::: Installing cray-play_xsvf
::: Removing unwanted files from the root

```

```

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.37336.662.

```

```

Running hssclone.
Image Clone Complete: /opt/cray/hss-images/image-7.2.0-1.0702.37336.662
Running hsspackage.
copying image
copying modules
running depmod
creating load file: /opt/cray/hss-images/default/HSS32.load
compressing initrd.img
Creating pxelinux.0 symlink
Running hssbootlink.
linking /opt/cray/hss-images/default/HSS32/bzImage-3.0.76-0.11.1_1.0702.8867-
cray_hss32 /opt/tftpboot/bzImage
linking /opt/cray/hss-images/default/HSS32/parameters /opt/tftpboot/
pxelinux.cfg/default
linking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img

```

2. Power down the system.

```
smw# xtcli power down s0
```

3. Reboot the cabinet controllers, then ensure that all cabinet controllers are up.

```
smw# xtccreboot -c all
xtccreboot: reboot sent to specified CCs
smw# xtalive -l cc
```

4. Power up the system.

```
smw# xtcli power up s0
```

Note that at this point the `xtcli` status output shows that all nodes are "off", because they have not yet been bounced.

5. Run the `xtdiscover` command to ensure that any changes made to the HSS database schema for new features are captured.

```
smw# xtdiscover
```

6. Exit from the `root` login.

```
smw# exit
```

7. Run the `rtr --discover` command to determine the exact configuration of the HSN.

```
crayadm@smw> rtr --discover
```

If the system was not bounced previously, the following message may be displayed:

```
System was not bounced in diagnostic mode, should I re-bounce? Continue (y/n)?
```

If so, respond with `y`.

8. Update the firmware. Execute the `xtzap` command to update the components.



CAUTION: The `xtzap` command is normally intended for use by Cray Service personnel only. Improper use of this restricted command can cause serious damage to the computer system.

```
crayadm@smw> xtzap -r -v s0
```

IMPORTANT: The Cray XC30 system also requires an update to the NVIDIA® BIOS (nvBIOS) for the NVIDIA K20X graphics processing units (GPUs). This update is done after CLE has been booted. For more information, see *CLE Installation and Configuration Guide (S-2444)*.

9. Use the output of the `xtzap` command to determine if any components need to be flashed.

While the `xtzap -a` command can be used to update all components with a single command, it may be faster to use the `xtzap -blade` command when only blade types need to be updated, or the `xtzap -t` command when only a single type needs to be updated. On larger systems, this can be a significant time savings.

This is the list of all cabinet level components:

```
cc_mc (CC Microcontroller)
cc_bios (CC Tolapai BIOS)
cc_fpga (CC FPGA)
chia_fpga (CHIA FPGA)
```

This is a list of all blade level components:

```
cbb_mc (CBB BC Microcontroller)
ibb_mc (IBB BC Microcontroller)
anc_mc (ANC BC Microcontroller)
bc_bios (BC Tolapai BIOS)
lod_fpga (LOD FPGA)
node_bios (Node BIOS)
loc_fpga (LOC FPGA)
qloc_fpga (QLOC FPGA)
```

If the output of the `xtzap` command shows that only a specific type needs to be updated, then use the `-t` option with that type (this example uses the `node_bios` type).

```
crayadm@smw> xtzap -t node_bios s0
```

If the output of the `xtzap` command shows that only blade component types need to be updated, then use the `-b` option:

```
crayadm@smw> xtzap -b s0
```

If the output of the `xtzap` command shows that both blade- and cabinet-level component types need to be updated, or if there is uncertainty about what needs to be updated, then use the `-a` option:

```
crayadm@smw> xtzap -a s0
```

- Execute the `xtzap -r -v s0` command again; all firmware revisions should report correctly, except `node_bios`; `node_bios` will display as "NOT_FOUND" until after the `xtbounce --linktune` command is run.

```
crayadm@smw> xtzap -r -v s0
```

- Execute the `xtbounce --linktune` command, which forces `xtbounce` to do full tuning on the system.

```
crayadm@smw> xtbounce --linktune=all s0
```

Continue with [Confirm the SMW is Communicating with System Hardware](#) on page 43.

Update SMW Software on the Boot Root and Shared Root

This procedure uses the `SMWinstallCLE` script to update the SMW software on the boot root and shared root for systems already running the Cray Linux Environment (CLE) software. The RPMs that `SMWinstallCLE` installs on the boot root and shared root will also be installed when `CLEinstall` runs during a CLE update.

TIP: Use this procedure only if the plan is to boot CLE after the SMW update but before updating the CLE software. Otherwise, if the plan is to update CLE software without booting CLE after the SMW update, it is safe to skip this procedure.

For more information about the `SMWinstallCLE` script, see the `SMWinstallCLE(8)` man page.

- As `root`, mount the release media by using one of the following commands, depending on the media type.
 - To install the update package from disk, place the `SMW 7.2UP04 Software DVD` in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as `/tmp/SMW_version` on the SMW and then substitute this path for `/media/cdrom` in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the `smw-image` ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

2. Update the `label_name` system set from the `/etc/sysset.conf` system set configuration file. In the following steps it is assumed that the label `label_name` is described in the `/etc/sysset.conf` system set configuration file. See the `sysset.conf(5)` man page for additional information about the `/etc/sysset.conf` file. For more detailed information about `SMWinstallCLE`, see the `SMWinstallCLE(8)` man page.

```
smw# /media/cdrom/utils/SMWinstallCLE --label=label_name
```

NOTE: The `SMWinstallCLE` command checks whether the boot node is booted. If it is booted, `SMWinstallCLE` prompts for confirmation that the system set being changed is not the one booted.

```
HH:MM:SS WARNING: Your bootnode is booted. Please confirm that the system
set you
intend to update is not booted.
Do you wish to proceed? [n] y
```

3. Unmount the SMW 7.2UP04 Software media.

```
smw# umount /media/cdrom
```

4. If using the update disk, eject the SMW 7.2UP04 Software DVD.

```
smw# eject
```

The SMW software is now updated. If the firewall is not yet configured, see [Set Up the SUSE Firewall and IP Tables](#) on page 45. Then continue to install the CLE software using [CLE Installation and Configuration Guide \(S-2444\)](#), which is provided with the CLE release package.

NOTE: To reconfigure a `LOGDISK`, `PMDISK`, or `DBDISK` when it is necessary to replace a failed drive with a new drive on a rack-mount SMW, see [Replace a Failed Disk Drive](#).

Configure the Simple Event Correlator (SEC)

The System Management Workstation (SMW) 7.2.UP04 release includes the Open Source simple event correlator (SEC) package, `sec-2.7.0`, and an SEC support package, `cray-sec-version`. The SEC support package contains control scripts to manage the starting and stopping of SEC around a Cray mainframe boot session, in addition to other utilities.

To use the Cray SEC, see [Configure SEC Software \(S-2542\)](#) for configuration procedures.

Finish Upgrading the SMW Software on the Active SMW

After updating the SMW software on the active SMW, reset the cluster resources.

1. Execute the `clean_resources` command on the active SMW (`smw1`).

```

smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

2. Wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -1` command.

Upgrade SMW Software on the Passive SMW

Before upgrading the SMW software on the passive SMW, use this procedure to prepare for the upgrade.

The examples in this procedure assume that `smw1` was the active SMW at the start of the upgrade.

1. Log into each SMW as `root`.
2. Bring the passive SMW online (take `smw2` out of standby mode).

```
smw2:~ # crm node online smw2
```

3. On the active SMW, force a failover to the passive SMW, then wait 30 seconds for the failover operation to complete.

```
smw1:~ # crm node standby
smw1:~ # sleep 30
```

NOTE: Ignore the failover errors. The failover operation will not complete successfully, because the second SMW has not been upgraded yet.

4. On the second SMW (`smw2`), turn on maintenance mode.

```
smw2:~ # crm configure property maintenance-mode=true 2> /dev/null
```

5. Determine the persistent device name of the `/home` directory on the boot RAID by displaying the configured device names.

```
smw2:~ # crm configure show | grep device | awk '{print $2 " " $3}' | sed 's/"//g'
```

```

device=/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9c01 directory=/home
device=/dev/disk/by-id/scsi-360080e500023bff6000006b3515d9bdf directory=/var/
lib/mysql
device=/dev/disk/by-id/scsi-360080e500023bff6000006b1515d9bc9 directory=/var/
opt/cray/disk/1

```

6. Mount the `/home` directory from the boot RAID.

NOTE: In the following command, replace `scsi-xxxxxxxx` with the the persistent device name from the previous step

```
smw2:~ # mount /dev/disk/by-id/scsi-xxxxxxxx /home
```

IMPORTANT: When you install the SMW software update in the next procedure, you must skip several steps on the second SMW.

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `phy7` and is slot 0, and the bootable backup disk is `phy6` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 266; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
    /dev/sda
    /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
    /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
    /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Physical slot 1:
```

```

/dev/sdc
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Physical slot 2:
/dev/sdd
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0
Physical slot 3:
/dev/sdb
/dev/disk/by-id/ata-ST9500620NS_9XF0665V
/dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0
Physical slot 4:
NOT INSTALLED
Physical slot 5:
NOT INSTALLED

```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```

smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD

```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and grub) drive names, the `device.map` mapping file used by grub should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```

# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical

```

```
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 263 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 263, the `fstab` lines would change from:

```
/dev/sda1 swap swap defaults 0 0
/dev/sda2 / ext3 acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 swap swap defaults 0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 / ext3 acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the device.map BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the grub bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the `grub` utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the `grub` utility and reinstall SMW root-drive boot blocks.

```

smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
      Checking if "/boot/grub/stage1" exists... yes
      Checking if "/boot/grub/stage2" exists... yes
      Checking if "/boot/grub/e2fs_stage1_5" exists... yes
      Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
      Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
grub> quit

```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

      Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
      Boot  Start      End    Blocks  Id System
           63  16771859   8385898+  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
      Boot  Start      End    Blocks  Id System
           * 16771860 312576704 147902422+  83  Linux

```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```

smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSs
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes

```

```
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080
```

```

          Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
          Boot  Start      End      Blocks  Id  System
swap / Solaris
          63  16771859      83828  82  Linux
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
          Boot  Start      End      Blocks  Id  System
          167719 312581807 156207044+ 83  Linux
```

```
Command (m for help): d
Partition number (1-4): 2
```

```
Command (m for help): d
Selected partition 1
```

```
Command (m for help): n
Command action
  e  extended
  p  primary partition (1-4)
```

```
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)
```

```
Command (m for help): n
Command action
  e  extended
  p  primary partition (1-4)
```

```
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807
```

```
Command (m for help): w
The partition table has been altered!
```

```
Calling ioctl() to re-read partition table.
Syncing disks.
```

- c. Display the boot backup disk partition layout and confirm it matches the phy7 sector information.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
```

4. Initialize the swap device.

```
smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
(DOS partition table detected). Use -f to force.
```

```
Setting up swapspace version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed
```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```
smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 37 mounts or
180 days, whichever comes first. Use tune2fs -c or -i to override.
```

6. Mount the new backup root file system on /mnt.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem            1K-blocks      Used Available Use% Mounted on
/dev/sda2              303528624    6438700 281671544   3% /
udev                  1030332        116  1030216   1% /dev
/dev/sdb2              306128812    195568 290505224   1% /mnt
```

The running root file system device is the one mounted on /.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
```

```
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the `root` and `swap` partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the grub boot utility. Within the grub boot utility:
1. Execute the `find` command to locate the drive designation that grub uses.
 2. Select the drive to which the boot blocks will be installed with the `root` command.
 3. Use the `setup` command to set up and install the grub boot blocks on that drive. The Linux grub utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the
  possible
  completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
grub> root (hd2,1)
root (hd2,1)
  Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
  embedded.
  succeeded
  Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1)/boot/grub/
  stage2 \
  /boot/grub/menu.lst"... succeeded
  Done.
grub> quit
```

IMPORTANT: For R815 SMWs, grub recreates `device.map` with the short names, not the persistent names. Do not trust the `/dev/sdx` names. Always use `find` when executing grub because it is possible that grub `root` may not be `hd2` the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

R630 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R630 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `pci-0000:03:00.0-scsi-0:0:0:0` and is slot 0, and the bootable backup disk is `pci-0000:03:00.0-scsi-0:0:1:0` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 273; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R630 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
  /dev/sda
  /dev/disk/by-id/scsi-35000c50079ab34b7
  /dev/disk/by-id/wwn-0x5000c50079ab34b7
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Physical slot 1:
  /dev/sdb
  /dev/disk/by-id/scsi-35000c50079ab71c4
  /dev/disk/by-id/wwn-0x5000c50079ab71c4
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
Physical slot 2:
  /dev/sdc
  /dev/disk/by-id/scsi-35000c50079ab313b
  /dev/disk/by-id/wwn-0x5000c50079ab313b
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
Physical slot 3:
  /dev/sdd
  /dev/disk/by-id/scsi-35000c50079ab4b4c
  /dev/disk/by-id/wwn-0x5000c50079ab4b4c
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
Physical slot 4:
  /dev/sde
  /dev/disk/by-id/scsi-35000c50079d05e70
  /dev/disk/by-id/wwn-0x5000c50079d05e70
  /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
Physical slot 5:
  NOT INSTALLED
Physical slot 6:
  NOT INSTALLED
Physical slot 7:
  NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and grub) drive names, the `device.map` mapping file used by grub should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name.

NOTE: `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r630 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
(hd1) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0
(hd2) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:2:0
(hd3) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:3:0
(hd4) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
(hd5) /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:5:0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 271 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 183, the `fstab` lines would change from:

```
/dev/sda1 swap          swap          defaults      0 0
/dev/sda2 /                ext3         acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1 swap swap
defaults 0 0
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 / ext3
acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the grub bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the grub utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the grub utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
  Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
      grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0
Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x000c3cc8
```

Blocks	Id	System	Device	Boot	Start	End
			/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part1		2048	67102719
33550336	82	Linux swap / Solaris				
			/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2	*	67102720	976773119
454835200	83	Linux				

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Partition number (1-4, default 1): 1
First sector (2048-976773167, default 2048): (Press the Enter key)
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-976773167, default 976773167): 67102719

Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4, default 2): 2
First sector (67102720-976773167, default 67102720): (Press the Enter key)
Using default value 67102720
Last sector, +sectors or +size{K,M,G} (67102720-976773167, default 976773167): (Press
the Enter key)
Using default value 976773167

Command (m for help): w
```

The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

- c. Display the boot backup disk partition layout and confirm it matches the pci-0000:03:00.0-scsi-0:0:0:0 sector information.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0

Disk /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0: 500.1 GB, 500107862016 bytes
255 heads, 63 sectors/track, 60801 cylinders, total 976773168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x7c334e96
```

Blocks	Id	System	Device	Boot	Start	End
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1	33550336	82	Linux swap		2048	67102719
/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2	454835224	83	Linux		67102720	976773167

4. Initialize the swap device.

```
smw# mkswap /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part1
Setting up swapspace version 1, size = 33550332 KiB
no label, UUID=8391498b-d159-469c-b766-66f00a28ff74
```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```
smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
28434432 inodes, 113708806 blocks
5685440 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
3471 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872, 71663616, 78675968,
    102400000

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 33 mounts or
180 days, whichever comes first. Use tune2fs -c or -i to override.
```

6. Mount the new backup root file system on /mnt.

```
smw# mount /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:1:0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem      1K-blocks    Used Available Use% Mounted on
/dev/sda2       447696736  9180648 437606420   3% /
udev            66029308     744  66028564   1% /dev
tmpfs           66029308    39540  65989768   1% /dev/shm
/dev/sdae       309637120 1107516 292800964   1% /var/opt/cray/disk/1
/dev/sdac       206424760 1963664 193975336   2% /home
/dev/sdad       154818540  474696 146479524   1% /var/lib/mysql
/dev/drbd_r0    961405840  247180 912322076   1% /var/lib/pgsql
/dev/sdb2       447696760  202940 424752060   1% /mnt
```

The running root file system device is the one mounted on /.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:0:0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 9129804 blocks.
DUMP: Volume 1 started with block 1 at: Wed Sep 16 15:43:08 2015
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
./tmp/rstmdir1442436041: (inode 27254928) not found on tape
./tmp/rstmode1442436041: (inode 27254931) not found on tape
DUMP: 77.64% done at 23626 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Wed Sep 16 15:50:09 2015
DUMP: Volume 1 9132800 blocks (8918.75MB)
DUMP: Volume 1 took 0:07:01
DUMP: Volume 1 transfer rate: 21693 kB/s
DUMP: 9132800 blocks (8918.75MB)
DUMP: finished in 421 seconds, throughput 21693 kBytes/sec
DUMP: Date of this level 0 dump: Wed Sep 16 15:40:41 2015
DUMP: Date this dump completed: Wed Sep 16 15:50:09 2015
DUMP: Average transfer rate: 21693 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the root and swap partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

```
smw# Vi /mnt/etc/fstab
```

For example, change

```
/dev/disk/by-id/scsi-35000c50079ab34b7-part1 swap      swap
defaults                                0 0
/dev/disk/by-id/scsi-35000c50079ab34b7-part2 /          ext3
acl,user_xattr                          1 1
```

to:

```
/dev/disk/by-id/scsi-35000c50079ab71c4-part1 swap      swap
defaults                0 0
/dev/disk/by-id/scsi-35000c50079ab71c4-part2 /          ext3
acl,user_xattr           1 1
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/scsi-35000c50079ab34b7-part2
```

with:

```
root=/dev/disk/by-id/scsi-35000c50079ab71c4-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_1
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
lists possible command completions. Anywhere else TAB lists the
possible
completions of a device/filename. ]
grub> find /THIS_IS_1
find /THIS_IS_1
(hd1,1)
grub> root (hd1,1)
root (hd1,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd1)
setup (hd1)
```

```

Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd1)"... 17 sectors are
embedded.
succeeded
Running "install /boot/grub/stage1 (hd1) (hd1)1+17 p (hd1,1)/boot/grub/
stage2 /boot/grub/menu.lst"... succeeded
Done.
grub> quit
quit

```

IMPORTANT: For R630 SMWs, grub recreates device.map with the short names, not the persistent names. Do not trust the /dev/sdx names. Always use find when executing grub because it is possible that grub root may not be hd2 the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

Update the SMW Software and Configuration

1. Open a terminal window, and su to root.

```
crayadm@smw> su - root
smw#
```

2. Mount the release media by using one of the following commands, depending on the media type.

- To install the update package from disk, place the SMW 7.2UP04 Software DVD in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as /tmp/SMW_version on the SMW and then substitute this path for /media/cdrom in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the smw-image ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

3. If postfix is not configured on the SMW, skip this step:

To prevent the master.cf and main.cf postfix configuration files from being recreated during software updates or fixes, ensure the following setting in the /etc/sysconfig/mail file on the SMW is set to "no":

```
MAIL_CREATE_CONFIG="no"
```

4. To see what SMW software will be updated in this new release, execute these commands prior to doing the update. This information is gathered, displayed, and contained in the log files during the SMWinstall process.
 - a. Check security and recommended updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GS
```

- b. Check Cray software updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GV
```

5. Create a new copy of the `SMWinstall.conf` configuration file and modify the new copy of the `SMWinstall.conf` file with site-specific requirements. Only `root` can modify the `SMWinstall.conf` configuration file. The `SMWinstall.conf` configuration file is created during the installation process by copying the `SMWinstall.conf` template from the distribution media. By default, the SMW configuration file is placed in `/home/crayadm/SMWinstall.conf`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
smw# chmod 644 /home/crayadm/SMWinstall.conf
smw# vi /home/crayadm/SMWinstall.conf
```

IMPORTANT: For an SMW HA system, define this variable as if the system was a standalone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.

For a description of the contents of the `SMWinstall.conf` file, see the `SMWinstall.conf(5)` man page.

6. Update the software with `SMWinstall`. `SMWinstall` checks for any inconsistency between the system and the `SMWinstall.conf` file settings, prompts for the root MySQL database password, and stores its log files in `/var/adm/cray/logs`.

```
smw# /media/cdrom/SMWinstall
...
```

```
Please enter your root DB password:
Please confirm your root DB password:
Password confirmed.
```

When `SMWinstall` finishes, it will suggest a reboot of the SMW.

7. If necessary, restore the locally modified versions of the following files.
 - a. If the installation site had locally modified versions of `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` before this SMW update, restore the local modifications to these files. During the upgrade procedure, the old files are saved in `/etc/syslog-ng/syslog-ng.conf-YYYYMMDDhhmm` and `/etc/rsyslog.conf.rpmsave`.
 - b. If the installation site had local modifications to `/opt/cray/hss/default/etc/auto.xtshutdown` before this SMW update, restore the local modifications to this file.
 - c. If the installation site had power management profiles, copy the files that were backed up from the `/opt/cray/hss/7.1.0/pm/profiles` directory into the new `/opt/cray/hss/default/pm/profiles` directory. See *Monitoring and Managing Power Consumption on the Cray XC30 System (S-0043)* for instructions on validating the restored profiles.
 - d. If the installation site had local modifications to `/opt/cray/hss/default/etc/sedc_srv.ini` before this SMW update, locate the destination of this symbolic link (see [Prepare to Upgrade or Update SMW Software](#) on page 245), compare the content of the local file to the distributed version of the file, and update the local file appropriately.
8. If the installation site has an `/opt/cray/hss/default/etc/xtdiscover.ini` file, the SMW update process does not overwrite an existing `xtdiscover.ini` file; the new version is named `xtdiscover.ini.dist`. Compare the

content of the new `xtdiscover.ini.dist` with the original `xtdiscover.ini` file, and update the `xtdiscover.ini` file appropriately.

NOTE: If the `xtdiscover.ini` file does not exist, then the `xtdiscover.ini.dist` file is copied to the `xtdiscover.ini` file.

9. Unmount the SMW 7.2UP04 Software media.

```
smw# umount /media/cdrom
```

10. If using the update disk, eject the SMW 7.2UP04 Software DVD.

```
smw# eject
```

11. Reboot the SMW.

```
smw# reboot
```

For a Cray XE or Cray XK system, continue to [Update the L0 and L1 Firmware](#).

Update the Controller Boot Image for the Passive SMW

Prerequisites

Before using this procedure, ensure that the SMW software and configuration has been updated on the passive SMW.

On the passive SMW, update only the controller boot image. Do not discover the hardware or update the firmware; these steps were done on the active SMW.

1. If necessary, log in as `root` to the passive SMW.

2. Update the controller boot image.

The version used in the command argument for `hss_make_default_initrd` should match that of the version specified in the `lsb-cray-hss` line in the output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Mon Sep 14 00:57:20 CDT 2015 on hssbld0 by bwdev
lsb-cray-hss-7.2.0-1.0702.37336.662
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.37336.662
::: Verifying base RPM list to the manifest
::: Installing filesystem hierarchy
::: Installing busybox
::: Installing base RPMs
::: Installing ssh
::: Removing /etc/securetty installed by the pam RPM
::: Installing sshfs
::: Installing rsh
::: Modifying /etc/pam.d/rlogin to remove securetty checking
::: Modifying /etc/pam.d/rsh to remove rhosts and nologin checking (Bug #779466)
::: Installing rsync
::: Installing atftp
::: Installing tcpdump
::: Installing ethtool
```

```

::: Installing syslog-ng
::: Installing logrotate
::: Installing ntp/ntpd
::: Installing strace
::: Installing screen
::: Installing minicom
::: Installing ppp
::: Installing mtd-utils
::: Installing /init
::: Installing file.rpm
::: Installing libgmodule
::: Installing Midnight Commander
::: Installing cray-viper
::: Installing spread
::: Installing coreboot-utils
::: Clearing init.d to be replaced by cray-hss32-filesystem
::: Creating initial etc files needed for root creation
::: Installing Cray kernel
::: Installing latest Cray kernel modules
::: Clearing select /boot items
::: Installing boot-parameters
::: Installing cray-hss32-scripts
::: Installing lsb-cray-hss-controllers
::: Installing cray-libconfig
::: Installing cray-bdm
::: Installing cray-play_xsvf
::: Removing unwanted files from the root

```

```

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.37336.662.

```

Running hssclone.

Image Clone Complete: /opt/cray/hss-images/image-7.2.0-1.0702.37336.662

Running hsspackage.

copying image

copying modules

running depmod

creating load file: /opt/cray/hss-images/default/HSS32.load

compressing initrd.img

Creating pxelinux.0 symlink

Running hssbootlink.

linking /opt/cray/hss-images/default/HSS32/bzImage-3.0.76-0.11.1_1.0702.8867-
cray_hss32 /opt/tftpboot/bzImage

linking /opt/cray/hss-images/default/HSS32/parameters /opt/tftpboot/
pxelinux.cfg/default

linking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img

Finish Upgrading the SMW Software on the Passive SMW

Prerequisites

Before beginning this procedure, ensure that the passive SMW has rebooted successfully at the end of the previous procedure.

After upgrading the SMW software on the passive SMW, make the following changes.

1. If there are local changes to `/etc/hosts` on `smw1`, manually copy `/etc/hosts` to `/etc/hosts` on `smw2`. The customized entries must be above the first section of "XT Cabinet x - y".

```
smw2:~ # cp /etc/hosts /etc/hosts.sav
smw2:~ # scp smw1:/etc/hosts /etc/hosts
```

Then edit the `/etc/hosts` file on `smw2` as follows:

- a. Change IP addresses `10.1.1.x`, `10.2.1.x`, `10.3.1.x`, and `10.4.1.x` to `10.1.1.y`, `10.2.1.y`, `10.3.1.y`, and `10.4.1.y` where if `x` is 2 `y` is 3 and if `x` is 3 `y` is 2.
- b. Change the line `smw1-ip smw1 smw1` to `smw2-ip smw2 smw2`.

2. On `smw2`, turn off maintenance mode.

```
smw2:~ # crm configure property maintenance-mode=false 2> /dev/null
```

3. Verify that the SMW HA services have started and that the second SMW is working properly.

```
smw2:~ # crm_mon -1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons   (lsb:rsm):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog (lsb:cray-syslog):             Started smw1
  homedir    (ocf::heartbeat:Filesystem):    Started smw1
  md-fs      (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs      (ocf::heartbeat:Filesystem):    Started smw1
  postgresql (lsb:postgresql):            Started smw1
  mysqld     (ocf::heartbeat:mysql):         Started smw1
```

NOTE: `crm_mon` may display different resource names, group names, or resource order on the system.

Finish the SMW Upgrade

Prerequisites

Before beginning this procedure, log into both SMWs as `root`.

Use this procedure to finish the SMW upgrade. The examples in this procedure assume that `smw1` was the active SMW at the start of the upgrade and is currently the passive SMW.

1. Reboot the currently passive SMW (`smw1`), if you did not reboot it after upgrading the SMW software on the passive SMW. Wait for the reboot to complete.
2. From the currently active SMW (`smw2`), bring the first SMW (`smw1`) online.

NOTE: Replace `smw1` with the host name of the SMW that was active at the start of the upgrade.

```
smw2:~ # crm node online smw1
```

3. On `smw1`, start file synchronizing.

```
smw1:~ # crm resource start fsync
```

4. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

5. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

6. Verify that all resources are running.
 - a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1    (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2    (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3    (ocf::heartbeat:IPaddr2):      Started smw1
```

```

ClusterIP4      (ocf::heartbeat:IPAddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):    Started smw1
Notification    (ocf::heartbeat:MailTo):              Started smw1
dhcpd           (lsb:dhcpd):                          Started smw1
fsync           (ocf::smw:fsync):                      Started smw1
hss-daemons     (lsb:rsmc):                            Started smw1
stonith-1       (stonith:external/ipmi):              Started smw2
stonith-2       (stonith:external/ipmi):              Started smw1
Resource Group: HSSGroup
  ml-fs         (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog   (lsb:cray-syslog):          Started smw1
  homedir       (ocf::heartbeat:Filesystem):    Started smw1
  md-fs         (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs         (ocf::heartbeat:Filesystem):    Started smw1
  postgresql    (lsb:postgresql):            Started smw1
  mysqld        (ocf::heartbeat:mysql):        Started smw1

```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```

smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

7. From either SMW, execute the `clear_failcounts` command to clean up any SMW HA resource errors.

```
smw1:~ # clear_failcounts
```

8. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, restart the VNC server.
9. Return the `postgresql` service to its pre-upgrade state (either on or off), as recorded in [Prepare the SMW HA System for an SMW Upgrade](#) on page 244.

```
smw1:~ # chkconfig postgresql state
```

Upgrade SMW HA Software

Prerequisites

Before you start this procedure:

- Ensure that both SMWs are running the required operating system and SMW software.
- Check that FN-6029 has been installed.

For more information, see [Before You Start an SMW HA Upgrade](#) on page 241.

To upgrade the SMW HA software, upgrade the active SMW first, then upgrade the other SMW.

1. Log in as `root` to the first SMW.

```
workstation> ssh root@smw1
```

2. In a separate terminal session, log in as `root` to the second SMW.

```
workstation> ssh root@smw2
```

3. Find the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -l | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

NOTE: The examples in this procedure assume that `smw1` is currently the active SMW.

4. Stop the OpenAIS service on both SMWs simultaneously.

```
smw1:~ # rcopenais stop
```

```
smw2:~ # rcopenais stop
```

5. Upgrade the SMW HA software on the active SMW (`smw1`).

- a. Mount the Cray SMW HA release media on the SMW.

- If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw1:~ # mount /dev/cdrom /media/cdrom
```

- If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

IMPORTANT: The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

The ISO file name depends on the release number and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For *path*, substitute the actual path to the ISO on the system.

```
smw1:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

- b. Navigate to the `/media/cdrom` directory.

```
smw1:~ # cd /media/cdrom
```

- c. Install the Cray SMW HA release software on the SMW.

```
smw1:~ # ./SMWHAinstall -v
```

- d. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.

- e. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw1:~ # cd
smw1:~ # umount /media/cdrom
smw1:~ # eject
```

6. Upgrade the SMW HA software on the other SMW (`smw2`).

- a. Mount the Cray SMW HA release media on the SMW.

- If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw2:~ # mount /dev/cdrom /media/cdrom
```

- If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

IMPORTANT:

The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

The ISO file name depends on the release number and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For *path*, substitute the actual path to the ISO on the system.

```
smw2:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

- b. Navigate to the `/media/cdrom` directory.

```
smw2:~ # cd /media/cdrom
```

- c. Install the Cray SMW HA release software on the SMW.

```
smw2:~ # ./SMWHAinstall -v
```

- d. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.
- e. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw2:~ # cd
smw2:~ # umount /media/cdrom
smw2:~ # eject
```

7. Start the OpenAIS service on both SMWs simultaneously.

```
smw1:~ # rcopenais start
```

```
smw2:~ # rcopenais start
```

8. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a ping command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

9. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a ping command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

10. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons   (lsb:rsms):                    Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
```

```
Resource Group: HSSGroup
ml-fs      (ocf::heartbeat:Filesystem): Started smw1
cray-syslog (lsb:cray-syslog):        Started smw1
homedir    (ocf::heartbeat:Filesystem): Started smw1
md-fs      (ocf::heartbeat:Filesystem): Started smw1
pm-fs      (ocf::heartbeat:Filesystem): Started smw1
postgresqld (lsb:postgresql):                Started smw1
mysqld     (ocf::heartbeat:mysql):    Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

11. Update the SMW HA cluster configuration.

- a. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, shut down the VNC server.

```
smw1:~ # /etc/init.d/vnc stop
```

- b. On the active SMW, change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

- c. Load the `ha-smw` module.

```
smw1:~ # module load ha-smw
```

- d. On the active SMW, execute the `SMWHAconfig` command with the `--update` option.

```
smw1:~ # ./SMWHAconfig --update
```

- e. When SMWHAconfig runs, it prompts for the virtual host name if the system is being updated from an older version of the release (such as from UP00 to UP01). Enter the virtual host name for the SMW HA cluster.
- f. If necessary, examine the log file. SMWHAconfig creates a log file in /opt/cray/ha-smw/default/hainst/SMWHAconfig.out.
- g. Reboot smw1 and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until smw1 has rejoined the cluster. After the SMW responds to a ping command, log into smw1, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that smw1 is online.

- h. Reboot smw2 and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until smw2 has rejoined the cluster. After the SMW responds to a ping command, log into smw2, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that smw2 is online.

- i. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                    Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs         (ocf::heartbeat:Filesystem):   Started smw1
  cray-syslog   (lsb:cray-syslog):            Started smw1
  homedir      (ocf::heartbeat:Filesystem):   Started smw1
  md-fs        (ocf::heartbeat:Filesystem):   Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):   Started smw1
  postgresql   (lsb:postgresql):            Started smw1
  mysqld       (ocf::heartbeat:mysql):       Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- j. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- k. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

- l. If necessary, restart the VNC server.
- 12.** If you have made local changes to the list of synchronized files and directories in `/etc/csync2/csync2_cray.cfg`, restore the local changes to the updated file.
- The installation procedure saves local changes in the file `/etc/csync2/csync2_cray.cfg.sav`. You must copy these changes into `csync2_cray.cfg`.
- a. On `smw1`, navigate to the `/etc/csync2` directory.
 - b. Edit the files `csync2_cray.cfg` and `csync2_cray.cfg.sav`.

NOTE: You can ignore the generic host entries near the top of the file. The `SMWHAconfig` command will restore site-specific host entries later in this procedure.
 - c. Locate the `group user_group` section in `csync2_cray.cfg.sav`, and copy the include and exclude lines into `csync2_cray.cfg`.
 - d. Save your changes to `csync2_cray.cfg` and exit the editor for both files.
- 13.** From either SMW, execute the `clear_failcounts` command to clean up any SMW HA resource errors.

```
smw1:~ # clear_failcounts
```

- 14.** Display the cluster status and verify that each resource has been started.

```
smw1:~ # crm_mon -1
Last updated: Mon Oct 27 01:19:23 2014
```

```

Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.

```

```
Online: [ smw1 smw2 ]
```

```

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons   (lsb:rsms):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog (lsb:cray-syslog):          Started smw1
  homedir     (ocf::heartbeat:Filesystem):    Started smw1
  md-fs       (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1

```

crm_mon may display different resource names, group names, or resource order on the system.

15. If any resource status is `Stopped`, execute the `clean_resources` command.

```
smw1:~ # clean_resources
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command.

16. Display the resource failcount data. All failcounts should be zero.

```

smw1:~# show_failcounts
node=smw1 scope=status name=fail-count-stonith-1 value=0
node=smw1 scope=status name=fail-count-stonith-2 value=0
node=smw1 scope=status name=fail-count-dhcpd value=0
node=smw1 scope=status name=fail-count-cray-syslog value=0
...

```

If there is a problem with the upgrade, see [Restore a Previous SMW HA Configuration](#) on page 399.

Upgrade CLE Software on the SMW HA System

When upgrading the SMW HA system, Cray recommends upgrading the CLE software after you have upgraded the operating system, SMW, and SMW HA software. If necessary, you can upgrade the CLE software before you upgrade the operating system, SMW, and SMW HA software.

Use the following procedures to upgrade the CLE software on the SMW HA system.

1. [Upgrade CLE Software on the Active SMW](#) on page 292
2. [Copy Boot Images and CLE Install Directory to the Passive SMW](#) on page 311
3. [Upgrade CLE Software on the Passive SMW](#) on page 311
4. [Finish the CLE Upgrade](#) on page 313

NOTE: The following procedures assume that `smw1` was the active SMW at the start of the upgrade.

Upgrade CLE Software on the Active SMW

Use the following procedures to upgrade the CLE software on the active SMW.

Before Starting the Update or Upgrade Process

Perform the following tasks before you install the CLE release package.

- Read the *README* file provided with the release for any installation-related requirements and corrections to this installation guide.
- Additional installation information may also be included in the following documents:
CLE 5.2.UP04 Release Errata, Limitations for CLE 5.2.UP04, Cray Linux Environment (CLE) Software Release Overview (S-2425), and Cray Linux Environment (CLE) Software Release Overview Supplement (S-2497).
- Verify that your System Management Workstation (SMW) is running Cray SMW Release 7.2.UP04 or later. You must install the SMW 7.2.UP04 release or later on your SMW before installing the CLE 5.2.UP04 release. If a specific SMW update package is required for your installation, that information is documented in the *README* file provided with the CLE 5.2.UP04 release. Type the following command to determine the HSS/SMW version:

```
crayadm@smw:~> cat /opt/cray/hss/default/etc/smw-release
7.2.UP04
```

Back Up the Current Software

Before you install the release package, back up the contents of the system set being updated or upgraded. Use the `xthotbackup` command to back up one system set to a second system set. For more information about using system sets, see [About System Set Configuration in /etc/sysset.conf](#) on page 88 and the `sysset.conf(5)` man page.

By default, `xthotbackup` copies only the boot node root and shared root file systems. Specify the `-a` option to copy all file systems in the system set (except for swap and Lustre) or specify the `-f` option to select a customized set of file system functions. The `-b` option makes the backup or destination system set bootable by changing the appropriate boot node and service node entries in `/etc/fstab`. Doing a live backup (`xthotbackup -L`) can reduce

the amount of time a CLE system is unavailable to the user community for the CLE backup and software upgrade process. For more information, see the `xthotbackup(8)` man page.

Back Up Current Software

Use the `xthotbackup` command to copy the disk partitions in one system set to a backup system set.



WARNING: If the source system set is booted, you should use the `xthotbackup -L` option. If not using the `xthotbackup -L` option, neither the source system set nor the destination system set should be used by a booted CLE system. Running `xthotbackup` with a booted system set or partition could cause data corruption.

1. If the Cray system is booted, use your site-specific procedures to shut down the system. For example, to shutdown using an automation file:

```
crayadm@smw:~> xtbootsys -s last -a auto.xtshutdown
```

For more information about using automation files see the `xtbootsys(8)` man page.

Although not the preferred method, alternatively execute these commands as `root` from the boot node to shutdown your system.

```
boot:~ # xtshutdown -y
boot:~ # shutdown -h now;exit
```

2. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

3. Run the `xthotbackup` command to copy from the source system set to the backup or destination system set. For example, if `BLUE` is the label for the source system set and `GREEN` is the label for the backup system set, execute the following command as `root`:

```
smw:~ # xthotbackup -a -b BLUE GREEN
```

NOTE: The `-a` option specifies all file system functions in the system set (except swap and Lustre). To specify a site-specific set of functions, use the `-f` option.

`xthotbackup` does not copy the swap partition for the boot node, however, if the `-b` option is specified, `mkswap` is invoked on the swap partition for the boot node in the destination system set to prepare a swap partition.

For more information, see the `xthotbackup(8)` man page.

Back Up Current Software Using `xthotbackup -L`

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Run the `xthotbackup` command to copy a booted system from the source system set to the backup or destination system set. For example, if `BLUE` is the label for the source system set and `GREEN` is the label for the backup system set, execute the following command:

```
smw:~ # xthotbackup -L -a -b BLUE GREEN
```

NOTE: The `-L` option will connect to the boot node (or other nodes that mount the file systems in the source system set) to backup the file systems in the destination system set. The `-a` option specifies all file system functions in the system set (except swap and Lustre). To specify a site-specific set of functions, use the `-f` option.

The `xthotbackup` command does not copy the swap partition for the boot node. However, if the `-b` option is specified, `mkswap` is invoked on the swap partition for the boot node in the destination system set to prepare a swap partition.

You are now ready to begin installing the software release package.

Upgrade CLE Software

A Cray system must be running CLE 4.0 or higher in order to upgrade the CLE software by using the procedures in this chapter.

Before You Begin

All upgrades, updates, and configuration changes are installed from the SMW to the `bootroot`, `sharedroot`, and (if applicable) the persistent `/var` file systems before booting the upgraded file systems. These file systems are mounted and modified during the procedure to install the release package.

An update or upgrade release package can be installed to an alternative root location if a system is configured to have more than one system set. A significant portion of the upgrade work can be done without using dedicated time if your Cray system is booted from a different system set. For example, if your system is running on the `BLUE` system set, and the `GREEN` system set is a backup of `BLUE`, then you can perform a large amount of the CLE upgrade process on the `GREEN` system set while the system is booted, thus reducing the amount of system downtime during upgrades/updates. These instructions will inform you when dedicated time is required.

The `/etc/sysset.conf` file describes which devices and disk partitions on the boot RAID are used for which system sets. For more information, see [About System Set Configuration in /etc/sysset.conf](#) on page 88 and the `sysset.conf(5)` man page.

If you are updating or upgrading a system set that is not running, you do not need to shut down your Cray system before you install the release package.



WARNING: If you are updating or upgrading a system set that is running, you must shut down your Cray system before installing the release package. For more information about system sets and system startup and shutdown procedures, see *Managing System Software for the Cray Linux Environment (S-2393)*.

If the persistent `/var` file system is shared between multiple system sets, you must verify that it is not mounted on the Cray system before you install the release package.

Install CLE Release Software on the SMW

Three DVDs are provided to install the CLE 5.2 release on a Cray system. The first is labeled `Cray CLE 5.2.UPnn Software` and contains software specific to Cray systems. Optionally, you may have an ISO image called `xc-sles11sp3-5.2.55d05.iso`, where `5.2.55` indicates the CLE release build level, and `d05` indicates the installer version.

To upgrade to the CLE 5.2 release from CLE or CLE 4.2 requires a second DVD labeled `Cray-CLEbase11sp3-yyyymmdd` and contains the CLE 5.2 base operating system, which is based on SLES 11 SP3. The third DVD is labeled `CentOS-6.5-x86_64-bin-DVD1.iso` and contains the CentOS 6.5 base operating system for CLE direct-attached Lustre (DAL) nodes.

Copy the Software to the SMW

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Mount the release media by using one of the following commands, depending on your media type.

If installing the release package from disk, place the Cray CLE 5.2.UPnn Software DVD in the CD/DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the release media using the ISO image, execute the following command, where `xc-sles11sp3-5.2.55d05.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro xc-sles11sp3-5.2.55d05.iso /media/cdrom
```

3. Copy all files to a directory on the SMW in `/home/crayadm/install.xtrel`, where `xtrel` is a site-determined name specific to the release being installed. For example:

```
smw:~# mkdir /home/crayadm/install.5.2.55
smw:~# cp -pr /media/cdrom/* /home/crayadm/install.5.2.55
```

4. Unmount the Cray CLE 5.2.UPnn Software media.

```
smw:~# umount /media/cdrom
```

5. For upgrading from CLE 5.1 or CLE 4.2 to CLE 5.2, you must mount the SLES 11 SP3 base media. Insert the Cray-CLEbase11sp3 DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the base operating system media using the ISO image, execute the following command, where `Cray-CLEbase11SP3-yyyymmdd.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro Cray-CLEbase11SP3-yyyymmdd.iso /media/cdrom
```

6. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` image.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Install CLE on the SMW

1. As `root`, execute the `CRAYCLEinstall.sh` installation script to upgrade the Cray CLE software on the SMW.

```
smw:~# /home/crayadm/install.5.2.56/CRAYCLEinstall.sh \
-m /home/crayadm/install.5.2.56 -u -v -w
```

2. At the prompt `Do you wish to continue?`, type **y** and press `Enter`.

The output of the installation script is displayed to the console. If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

Prepare the Configuration for Software Installation

You may need to update the `CLEinstall.conf` configuration file. The `CLEinstall.conf` file that was created during the first installation of this system can be used during an installation to the alternative root location. For a description of the contents of this file, see [About Installation Configuration Files](#) on page 74 or the `CLEinstall.conf(5)` man page.

Based on the settings you choose in the `CLEinstall.conf` file, the `CLEinstall` program then updates other configuration files. A template `CLEinstall.conf` is provided on the distribution media. Your site-specific copy is located in the installation directory from the previous installation; for example `/home/crayadm/install.5.2.14/CLEinstall.conf`.



WARNING: Any configuration data which is in `CLEinstall.conf` that was manually changed on a system after the last software update must be kept up to date before running `CLEinstall` for an upgrade or an update. Doing so will prevent spending much time tracking down problems that could have been avoided.

During update and upgrade installations, the `/opt/cray/hss/default/etc/auto.xtshutdown` automated shut-down file is overwritten by the newer shut-down file that corresponds to the update/upgrade release. The old shut-down file will be saved as `/opt/cray/hss/default/etc/auto.xtshutdown.rpmsave`. If your site has made local changes to the autofile, you will need to review the changes and reapply them to the new file following the update or upgrade.

NOTE: If problems with the hosts file are detected after the update or upgrade, you may need to use the copies of `/etc/hosts` that `CLEinstall` saves on bootroot and `/opt/xt-images/templates/default/etc` with `hosts.preinstall.$$` and `hosts.postinstall.$$`.

Prepare the `CLEinstall.conf` Configuration File

1. If you have an existing `CLEinstall.conf` file, use the `diff` command to compare it to the template in `/home/crayadm/install.xtre1`. For example:

```
smw:~# diff /home/crayadm/install.5.2.56/CLEinstall.conf \
/home/crayadm/install.5.2.14/CLEinstall.conf
21c21
< xthostname=mycray
---
> xthostname=crayhostname
24c24
< node_class_login_hostname=mycray
---
> node_class_login_hostname=crayhostname
smw:~#
```

NOTE: The `CLEinstall` program generates INFO messages suggesting that you remove deprecated parameters from your local `CLEinstall.conf` file.

2. Edit the `CLEinstall.conf` file in the temporary directory `/home/crayadm/install.xtre1` and make necessary changes to enable any new features you are configuring for the first time with this system software upgrade.

NOTE: The CLEinstall program checks that the `/etc/opt/cray/sdb/node_classes` file and the `node_class[*]` parameters in `CLEinstall.conf` agree. If you made changes to `/etc/opt/cray/sdb/node_classes` since your last CLE software installation or upgrade, make the same changes to `CLEinstall.conf`.

```
smw:~# cp -p /home/crayadm/install.5.2.56/CLEinstall.conf \
/home/crayadm/install.5.2.56/CLEinstall.conf.save
smw:~# chmod 644 /home/crayadm/install.5.2.56/CLEinstall.conf
smw:~# vi /home/crayadm/install.5.2.56/CLEinstall.conf
```

For a complete description of the contents of this file, see [About Installation Configuration Files](#) on page 74.

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

Run the CLEinstall Installation Program

The CLEinstall installation program upgrades the CLE software for your configuration by using information in the `CLEinstall.conf` and `sysset.conf` configuration files.

IMPORTANT: CLEinstall modifies Cray system entries in `/etc/hosts` each time you update or upgrade your CLE software. For additional information, see [Maintain Node Class Settings and Hostname Aliases](#) on page 76.

If the update or upgrade you are applying modifies configuration information in the `alps.conf` file, your existing `alps.conf` parameters will be automatically merged into the new file and your original file will be saved (in the same directory) as `alps.conf.unmerged`. If you experience problems with ALPS immediately following an update or upgrade, you can replace `alps.conf` with `alps.conf.unmerged` and execute `/etc/init.d/alps restart` on the boot and SDB nodes to restore your original configuration.

During a CLE update or upgrade, CLEinstall disables the execution bits of all scripts in the `/etc/cron.hourly`, `/etc/cron.daily`, `/etc/cron.weekly`, and `/etc/cron.monthly` directories of the `bootroot` and default view of the shared root with a `chmod ugo-x` command. If there are site-specific cron scripts in these directories, you will need to re-enable the execute permission on them after performing a CLE update or upgrade. Any scripts in these directories which have been node-specialized or class-specialized via `xtopview` will not be changed by the CLE update or upgrade. Only the `bootroot` and the default view of the shared root will be modified.

The following CLEinstall options are required or recommended for this type of installation:

`--upgrade`

Specify that this is an update or upgrade rather than a full system installation.

`--label=system_set_label`

Specify the system set that you are using to install the release.

`--XTrelease=release_number`

Specify the target CLE release and build level that you are upgrading to, for example 5.2.55.

`--CLEmedia=directory`

Specify the directory on the SMW where you copied the CLE software media. For example, `/home/crayadm/install.release_number`.

`--configfile=CLEinstall_configuration_file`

Specify the path to the `CLEinstall.conf` file that you edited in [Preparing the CLEinstall.conf configuration file](#).

`--Basemedia=directory`

Specify which directory the CLE base operating system media is mounted on. The default is `/media/cdrom`.

```
--Centosmedia=directory
```

Specify the directory where the CentOS software media has been mounted. The `--Centosmedia` option is required when installing or upgrading CLE with direct-attached Lustre (DAL). For example, the CentOS image mount point could be `/media/Centosbase`.

For a full description of the `CLEinstall` command options and arguments, see [Run the CLEinstall Program](#) on page 94 or the `CLEinstall(8)` man page.

Run CLEinstall

1. Invoke the `CLEinstall` program on the SMW. `CLEinstall` is located in the directory you created in [Copy the Software to the SMW](#).

```
smw:~# /home/crayadm/install.5.2.55/CLEinstall --upgrade \  
--label=system_set_label --XTrelease=5.2.55 \  
--configfile=/home/crayadm/install.5.2.55/CLEinstall.conf \  
--CLEmedia=/home/crayadm/install.5.2.55 \  
--Basemedia=/media/cdrom
```

When DAL is enabled, the `CLEinstall` program requires the `--Centosmedia` option.

```
smw:~# /home/crayadm/install.5.2.55/CLEinstall --upgrade \  
--label=system_set_label --XTrelease=5.2.55 \  
--configfile=/home/crayadm/install.5.2.55/CLEinstall.conf \  
--CLEmedia=/home/crayadm/install.5.2.55 \  
--Basemedia=/media/cdrom --Centosmedia=/media/Centosbase
```

2. Examine the initial messages directed to standard output. Log files are created in `/var/adm/cray/logs` and named by using a timestamp that indicates when the install script began executing. For example:

```
08:57:48 Installation output will be captured in /var/adm/cray/logs/  
CLEinstall.p3.20140911085748.CLE52-P3.stdout.log  
08:57:48 Installation errors (stderr) will be captured in /var/adm/cray/logs/  
CLEinstall.p3.20140911085748.CLE52-P3.stderr.log  
08:57:48 Installation debugging messages will be captured in /var/adm/cray/logs/  
CLEinstall.p3.20140911085748.CLE52-P3.debug.log
```

The naming conventions of these logs are:

```
CLEinstall.p#.YYYYMMDDhhmmss.$LABEL.logtype.log
```

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format.

`$LABEL` is the system set label (in the example above, `CLE52-P3`).

`logtype` is `stdout` (standard output), `stderr` (standard error), or `debug`.

Also, log files are created in `/var/adm/cray/logs` each time `CLEinstall` calls `CRAYCLEinstall.sh`. For example:

```
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.01-B.log  
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.02-B.log  
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.03-B.log
```

```

.
.
.
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.17-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.18-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.19-S.log

```

The naming conventions of these logs are:

`CRAYCLEinstall.sh.p#.YYYYMMDDhhmmss.$LABEL.sequence#-root.log`

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format. This is the same timestamp used for the log files of the `CLEinstall` program instance that called `CRAYCLEinstall.sh`.

`$LABEL` is the system set label.

`sequence#` is an increasing count that specifies each invocation of `CRAYCLEinstall.sh` by `CLEinstall`.

`root` is either `B` (bootroot) or `S` (sharedroot), specifying the root modified by the `CRAYCLEinstall.sh` call.

3. `CLEinstall` validates `sysset.conf` and `CLEinstall.conf` configuration settings and then confirms the expected status of your boot node and file systems.

Confirm that the installation is proceeding as expected, respond to warnings and prompts, and resolve any issues. For example:

- If you are installing to a system set that is not running, and you did not shut down your Cray system, respond to the following warning and prompt:

```

WARNING: Your bootnode is booted. Please confirm that the
system set you are intending to update is not booted.
Do you wish to proceed?[n]:

```



WARNING: If the boot node has a file system mounted and `CLEinstall` on the SMW creates a new file system on that disk partition, the running system will be corrupted.

- If you have configured file systems that are shared between two system sets, respond to the following prompt to confirm creation of new file systems:

```

09:21:24 INFO: The PERSISTENT_VAR disk function for the LABEL system set is marked shared.
09:21:24 INFO: The /dev/sdrl disk partition will be mounted on the SMW for PERSISTENT_VAR
disk function. Confirm that it is not mounted on any nodes in a running XT
system before continuing.
Do you wish to proceed?[n]:y

```

- If the `node_class_idx` parameters do not match the existing `/etc/opt/cray/sdb/node_classes` file, you are asked to confirm that your hardware configuration has changed. If your hardware has not changed, abort `CLEinstall` and correct the node class configuration in `CLEinstall.conf` and/or the `node_classes` file. Respond to the following warning and prompt:

```

09:21:41 INFO: There are 5 WARNINGS about discrepancies between CLEinstall.conf
and /etc/opt/cray/sdb/node_classes
09:21:41 INFO: If you ARE adding service nodes, then you may proceed and CLEinstall
will adjust the /etc/opt/cray/sdb/node_classes file to match the setting
s in CLEinstall.conf and may remove some node-specialized files from the shared
root specialized /etc.
09:21:41 INFO: If you ARE NOT adding service nodes, then stop CLEinstall now
to correct the problem.
Do you wish to proceed?[n]:

```

CLEinstall may resolve some issues after you indicate that you want to proceed; for example, disk devices are already mounted, boot image file or links already exist, HSS daemons are stopped on the SMW.



CAUTION: Some problems can be resolved only through manual intervention via another terminal window or by rebooting the SMW; for example, a process is using a mounted disk partition, preventing CLEinstall from unmounting the partition.

4. Monitor the debug output. Create another terminal window and invoke the `tail` command by using the path and timestamp displayed when CLEinstall was run.

```
smw~:# tail -f /var/adm/cray/logs/CLEinstall.p#. YYYYMMDDhhmmss. $LABEL.debug.log
```

5. Locate the following warning and prompt in the CLEinstall console window and type `y`.

```
*** Preparing to UPGRADE software on system set label system_set_label. Do you
wish to proceed? [n]
```

The CLEinstall program now installs the release software. This command runs for 30 minutes or more for updates and 90 minutes for an upgrade, depending on your system configuration.

6. Monitor the output to ensure that your installation is proceeding without error. Several error messages from the `tar` command are displayed as the persistent `/var` is updated for each service node. You may safely ignore these messages.
7. Confirm that the CLEinstall program has completed successfully.

On completion, the CLEinstall program generates a list of command hints to be run as the next steps in the update or upgrade process. These commands are customized, based on the variables in the `CLEinstall.conf` and `sysset.conf` files, and include runtime variables such as PID numbers in file names. The list of command hints is written to the `CLEinstall.command_hints.timestamp` file in the installer log directory.

Complete the upgrade/update and configuration of your Cray system by using both the commands that the CLEinstall program provides and the information in the remaining sections of this chapter.

As you complete these procedures, you can cut and paste the suggested commands from the output window or from the window created in a previous step that tailed the debug file. The log files created in `/var/adm/cray/logs` for `CLEinstall.P#. YYYYMMDDhhmmss. $LABEL.stdout.log` and `CLEinstall.P#. YYYYMMDDhhmmss. $LABEL.debug.log` also contain the suggested commands.

A CLE upgrade requires corrective boots of the system, which are contained in the command hints. The command hints contain commands for the following tasks:

1. Boot the boot node and SDB node (all upgrades).
2. Generate the new ECDSA SSH host key because of the upgrade to SLES 11 SP3 (all upgrades).
3. Update the SDB schema for upgrades from older CLE versions, which is described in [Upgrade the SDB Database Utilities with a CLE Update Package](#) (some upgrades).
4. Shut down the boot and SDB nodes, and boot the full system (all upgrades).
5. Run `shell_ssh.sh` once the service nodes are booted up.

When upgrading DAL, the command hints contain instructions for building the DAL image, creating the IMPS config set, provisioning the DAL image, and further Lustre configuration tasks. The details for these DAL tasks are described in [Upgrade DAL on XE Systems](#).

Create Boot Images

The Cray CNL compute nodes and Cray service nodes use RAM disks for booting. Service nodes and CNL compute nodes use the same `initramfs` format and workspace environment. This space is created in `/opt/xt-images/machine-xtrelease-LABEL-partition/nodetype`, where `machine` is the Cray hostname, `xtrelease` is the build level for the CLE release, `LABEL` is the system set label used from `/etc/sysset.conf`, `partition` describes either the full machine or a system partition, and `nodetype` is either `compute` or `service`.



CAUTION: Existing files in `/opt/xt-images/templates/default` are copied into the new bootimage work space. In most cases, you can use the older version of the files with the upgraded system. However, some file content may have changed with the new release. Verify that site-specific modifications are compatible. For example, use existing copies of `/etc/hosts`, `/etc/passwd` and `/etc/modprobe.conf`, but if `/init` changed for the template, the site-modified version that is copied and used for CLE 5.2 may cause a boot failure.

Follow the procedures in this section to prepare the work space in `/opt/xt-images`. For more information about configuring boot images for service and compute nodes, see the `xtclone(8)` and `xtpackage(8)` man pages.

Prepare Compute and Service Node Boot Images

The `shell_bootimage_LABEL.sh` script prepares boot images for the system set specified by `LABEL`. For example, if your system set has the label `BLUE` in `/etc/sysset.conf`, invoke `shell_bootimage_BLUE.sh` to prepare a boot image. This script uses `xtclone` and `xtpackage` to prepare the work space in `/opt/xt-images`.

NOTE: When upgrading a system using direct-attached Lustre (DAL), use the `-d` option. This option specifies that the CentOS DAL be included. For more information, see [Create a Boot Image That Includes the DAL Image](#).

For more information about configuring boot images for service and compute nodes, see the `xtclone(8)` and `xtpackage(8)` man pages.

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Run the `shell_bootimage_*.sh` script, where `LABEL` is the system set label specified in `/etc/sysset.conf` for this boot image.

Specify the `-c` option to automatically create and set the boot image for the next boot. For example, if the system set label is `BLUE`:

```
smw:~# /var/opt/cray/install/shell_bootimage_BLUE.sh -c
```

For information about additional options accepted by this script, use the `-h` option to display a help message.

Enable Boot Node Failover

NOTE: Boot node failover is an optional CLE feature.

If boot-node failover has been configured for the first time, follow these steps. If boot-node failover has not been configured, skip this procedure.

To enable bootnode failover, you must set `bootnode_failover` parameters in the `CLEinstall.conf` file before you run the `CLEinstall` program. For more information, see [Configure Boot Node Failover](#) on page 81.

In this example, the primary boot node is `c0-0c0s0n1` (`node_boot_primary=1`) and the backup or alternate boot node is `c0-0c1s1n1` (`node_boot_alternate=61`).

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

1. As `crayadm` on the SMW, halt the primary and alternate boot nodes.



WARNING: Verify that the system is shut down before you invoke the `xtcli halt` command.

```
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
```

2. Specify the primary and backup boot nodes in the boot configuration.

If the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the boot node for the entire system.

```
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command to select the boot node for the designated partition.

```
crayadm@smw:~> xtcli part_cfg update pN -b c0-0c0s0n1,c0-0c1s1n1
```

3. To use boot-node failover, enable the STONITH capability on the blade or module of the primary boot node. Use the `xtdaemonconfig` command to determine the current STONITH setting.

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary boot node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

4. To enable STONITH on the primary boot node blade, type the following command:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 stonith=true
c0-0c0s0: stonith=true
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.

Enable SDB Node Failover

NOTE: SDB node failover is an optional CLE feature.

If SDB node failover has been configured for the first time, follow these steps. If SDB node failover has not been configured, skip this procedure.

In addition to this procedure, refer to [Configure Boot Automation for SDB Node Failover](#) on page 140 after you have completed the remaining configuration steps and have booted and tested your system.

To enable SDB node failover, you must set `sdbnode_failover` parameters in the `CLEinstall.conf` file before you run the `CLEinstall` program. For more information, see [Configure SDB Node Failover](#) on page 82.

In this example, the primary SDB node is `c0-0c0s2n1` (`node_sdb_primary=5`) and the backup or alternate SDB node is `c0-0c1s3n1` (`node_sdb_alternate=57`).

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

1. Invoke `xtdaemonconfig` to determine the current STONITH setting on the blade or module of the primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary SDB node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

2. Enable STONITH on your primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s2 stonith=true
c0-0c0s2: stonith=true
The expected response was received.
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.

3. Specify the primary and backup SDB nodes in the boot configuration.

For example, if the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the primary and backup SDB nodes.

```
crayadm@smw:~> xtcli halt c0-0c0s2n1,c0-0c1s3n1
crayadm@smw:~> xtcli boot_cfg update -d c0-0c0s2n1,c0-0c1s3n1
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command:

```
crayadm@smw:~> xtcli part_cfg update pN -d c0-0c0s2n1,c0-0c1s3n1
```

Run Post-CLEinstall Commands

1. Unmount and eject the release software DVD from the SMW DVD drive if it is still loaded.

```
smw:~# umount /media/cdrom
smw:~# umount /media/Centosbase
smw:~# eject
```

2. Run the `shell_post_install.sh` script on the SMW to unmount the boot root and shared root file systems and perform other cleanup as needed.

```
smw:~# /var/opt/cray/install/shell_post_install.sh /bootroot0 /sharedroot0
```



WARNING: Exercise care when you mount and unmount file systems. If you mount a file system on the SMW and boot node simultaneously, you may corrupt the file system.

3. Confirm that the `shell_post_install.sh` script successfully unmounted the boot root and shared root file systems.

If a file system does not unmount successfully, the script displays information about open files and associated processes (by using the `lsof` and `fuser` commands). Attempt to terminate processes with open files and if necessary, reboot the SMW to resolve the problem.

Update the SDB Database Schema

Upgrading to CLE 5.2 may require that the Service Database (SDB) database schema be upgraded. The command hints indicate whether this is necessary. Refer to [Upgrade the SDB Database Utilities with a CLE Update Package](#) before continuing.

Configure Optional Services

If you enabled an optional service you were not previously using in [Prepare the CLEinstall.conf Configuration File](#) on page 296, you may need to perform additional configuration steps. Follow the procedures in the appropriate optional section in [Install CLE on a New System](#) on page 91 or in [Managing System Software for the Cray Linux Environment \(S-2393\)](#).

If you configured an optional CLE feature or service during a previous installation or upgrade, no additional steps are required.

Configure MAMU Nodes

1. On the boot node, run this script, which ensures the keys are correct on the MAMU nodes (the `postproc` node class in this and previous examples).

```
boot# /var/opt/cray/install/shell_ssh.sh
```

2. Modify the `sshd_config` and `/etc/fstab` files for the new `postproc` class.

```
boot# xtopview -c postproc -m "setting up postproc nodes"
class/postproc:/# xtspec /etc/fstab
```

Add this line:

```
ufs: /ufs/home          /ufs/home      nfs      tcp,rw  0 0
```

```
class/postproc:/ # xtspec /etc/ssh/sshd_config
class/postproc:/ # vi /etc/ssh/sshd_config
```

Strip any `MatchUser` blocks from the bottom of the `sshd_config` file. Save and close the file.

3. Run these commands to restrict logins on the `postproc` nodes to only the `crayadm` administrative account and `root`, which is necessary to provide out of memory protection.

```
class/postproc:/ # xtspec /etc/ssh/sshd_config
class/postproc:/ # echo "AllowUsers root crayadm" >> /etc/ssh/sshd_config
class/postproc:/ # exit
```

Boot and Test the System

IMPORTANT: If you configured optional services for the first time during this upgrade and deferred updating the boot image, update the boot image now by following [Prepare Compute and Service Node Boot Images](#) on page 374.

Your system is now upgraded. Boot the system using either `xtbootsys` interactive mode or a boot automation file.

Boot the System with Interactive `xtbootsys`

1. Boot the boot node, followed by the SDB node, and then all remaining service nodes, but do not boot the CNL compute nodes.

```
crayadm@smw> xtbootsys
```

2. Update the SSH known host keys for `root@boot` by running this script on the boot node after all of the service nodes complete their boot.

```
smw# ssh root@boot
boot# /var/opt/cray/install/shell_ssh.sh
```

3. Boot the CNL compute nodes one of these ways.

- a. Use menu option 17 in `xtbootsys`.
- b. Execute this command.

```
crayadm@smw> xtcli -s boot CNL0 -o compute s0
```

Boot the System with a Boot Automation File

1. Merge the old boot automation file with `/opt/cray/hss/default/etc/auto.generic.cnl` to create a new boot automation file, `/opt/cray/hss/default/etc/auto.mycray`.
2. Boot the CLE system using a boot automation file.

```
crayadm@smw> xtbootsys -a auto.mycray
```

3. Update the SSH known host keys for `root@boot` by running this script on the boot node after all of the service nodes complete their boot.

```
smw# ssh root@boot
boot# /var/opt/cray/install/shell_ssh.sh
```

Flash the nvBIOS for Kepler GPUs

A Cray XC30 system with NVIDIA® Tesla® SXM modules requires an update to the NVIDIA BIOS (nvBIOS) for the NVIDIA K20X and K40s graphics processing units (GPUs). The nvBIOS is unique for each SXM-1 Kepler™ SKU, based on the type of heat sink, as shown below.

GPU Type	Board SKU	Production Firmware Image Version
Kepler K20X (13 fin)	P2085 SKU 202	80.10.44.00.02

GPU Type	Board SKU	Production Firmware Image Version
Kepler K20X (20 fin)	P2085 SKU 212	80.10.44.00.04
Kepler K20X (30 fin)	P2085 SKU 222	80.10.44.00.05
Kepler K40s (13 fin)	P2085 SKU 209	80.80.4B.00.03
Kepler K40s (20 fin)	P2085 SKU 219	80.80.4B.00.04
Kepler K40s (30 fin)	P2085 SKU 229	80.80.4B.00.05

The CLE software includes a script that automatically determines the SKU version and flashes the nvBIOS with the appropriate firmware.

TIP: You can use the `cselect` command to identify the number and location of the Kepler GPUs. This example shows a system with K20X GPUs on four nodes.

```
login:~# cselect -c "subtype.eq.'nVidia_Kepler'"
4
login:~# cselect -e "subtype.eq.'nVidia_Kepler'"
70-73
```

1. As root on the login node, set the allocation mode for all compute nodes to interactive.

```
login:~# xtprocadm -km interactive
```

2. Change to the directory containing the NVIDIA scripts.

```
login:~# cd /opt/cray/cray-nvidia/default/bin
```

3. To flash the Kepler K20X GPUs, for example, choose one of the following options.

- To update the entire system:

```
login:~# aprun -n `cselect -c "subtype.eq.'nVidia_Kepler'"` \
-N 1 -L `cselect -e "subtype.eq.'nVidia_Kepler'"` \
./nvFlashBySKU -b
```

- To update a single node or set of nodes:

```
login:~# aprun -n PEs -N 1 -L node_list ./nvFlashBySKU -b
```

NOTE: Replace *PEs* with the number of nodes; replace *node_list* with a comma-separated list of NIDs.

For example, to flash four GPUs on nodes 70-73:

```
login:~# aprun -n 4 -N 1 -L 70,71,72,73 ./nvFlashBySKU -b
c0-0c0s1n0: Nid 70: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n1: Nid 71: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n2: Nid 72: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n3: Nid 73: Successful Cray Graphite K20X nvBIOS flash
```

4. If there is a flash failure, `nvFlashBySKU` displays an error message with the failing node ID, as in this example:

```
c0-0c0s1n3: Nid 73: Failed Cray Graphite K20X nvBIOS flash
```

Depending on the type of failure, nvFlashBySKU might display additional information, if available. No flashing is done on unsupported SKUs.

If a GPU fails to flash, the SXM-1 card must be replaced.

- After flashing is successful, use `xtbootstys` to reboot the nodes from the SMW. For example:

```
crayadm@smw:~> xtbootstys --reboot -L CNL0 -r "rebooting after nvBIOS update" \
c0-0c0s1n0,c0-0c0s1n1,c0-0c0s1n2,c0-0c0s1n3
```

TIP: You can use `xtprocdadmin` on the login node to determine each node name from the `cnsselect` output, as in this example:

```
login:~# xtprocdadmin -n `cnsselect -e "subtype.eq.'nVidia_Kepler'"`
  NID      (HEX)      NODENAME      TYPE      STATUS      MODE
   70      0xf8      c0-0c0s1n0    compute    up          batch
   71      0xf9      c0-0c0s1n1    compute    up          batch
   72      0xfa      c0-0c0s1n2    compute    up          batch
   73      0xfb      c0-0c0s1n3    compute    up          batch
```

- After the reboot is successful, log on to the login node as root and change to the directory containing the NVIDIA scripts.

```
login:~# cd /opt/cray/cray-nvidia/default/bin
```

To verify that all GPUs are reporting the correct nvBIOS version (see the table above), choose one of the following options:

- To display the nvBIOS versions for the entire system:

```
login:~# aprun -n `cnsselect -c "subtype.eq.'nVidia_Kepler'"` \
-N 1 -L `cnsselect -e "subtype.eq.'nVidia_Kepler'"`\
./xkcheck -n -c -f | grep Version
```

- To display the nvBIOS versions for a single node or set of nodes:

```
login:~# aprun -n PEs -N 1 -L node_list \
./xkcheck -n -c -f | grep Version
```

Replace *PEs* with the number of nodes; replace *node_list* with a comma-separated list of NIDs.

For example:

```
login:~# aprun -n 4 -N 1 -L 70,71,72,73 \
./xkcheck -n -c -f | grep Version
4 nodes report VBIOS Version           : 80.10.3D.00.05
```

- Reset the compute nodes to the normal batch or interactive mode using the `xtprocdadmin` command.

Test the System for Basic Functionality

- If the system was shut down by using `xtshutdown`, remove the `/etc/nologin` file from all service nodes to permit a non-root account to log on.

```
smw:~# ssh root@boot
boot:~ # xtunspec -r /rr/current -d /etc/nologin
```

- Log on to the login node as `crayadm`.


```

W  waiting or non-running job           X  down compute node
Y  down or admindown service node      Z  admindown compute node

Available compute nodes:                0 interactive,          650 batch

```

The `xtprocdadmin` command displays the current values of processor flags and node attributes. The output for Cray XE and Cray XK systems follows.

```

crayadm@login:~> xtprocdadmin
  NID  (HEX)  NODENAME  TYPE  STATUS  MODE
    0   0x0   c0-0c0s0n0  service  up  interactive
    2   0x2   c0-0c0s1n0  service  up  interactive
    4   0x4   c0-0c0s2n0  service  up  interactive
    6   0x6   c0-0c0s3n0  service  up  interactive
. . .
   93  0x5d   c0-0c2s1n3  service  up  interactive
   94  0x5e   c0-0c2s0n2  service  up  interactive
   95  0x5f   c0-0c2s0n3  service  up  interactive

```

The output for Cray XC30 systems follows.

```

crayadm@login:~> xtprocdadmin
  NID  (HEX)  NODENAME  TYPE  STATUS  MODE
    1   0x1   c0-0c0s0n1  service  up  batch
    2   0x2   c0-0c0s0n2  service  up  batch
    5   0x5   c0-0c0s1n1  service  up  batch
    6   0x6   c0-0c0s1n2  service  up  batch
    8   0x8   c0-0c0s2n0  compute  up  batch
    9   0x9   c0-0c0s2n1  compute  up  batch
   10  0xa   c0-0c0s2n2  compute  up  batch

```

The `apstat` command displays the current status of all applications running on the system.

```

crayadm@login:~> apstat -v
Compute node summary
  arch  config  up  resv  use  avail  down
  XT    733    733  107   89   626    0

Total pending applications: 4
Pending Pid  User  w:d:N  NID  Age  Command  Why
17278  crayadm  1848:1:24  5  0h53m  ./app1  Busy
17340  crayadm  1848:1:24  5  0h53m  ./app1  Busy
17469  crayadm  1848:1:24  5  0h52m  ./app1  Busy
26155  crayadm  1848:1:24  5  0h12m  ./app2  Busy

Total placed applications: 2
  Apid  ResId  User  PEs  Nodes  Age  State  Command
1631095  135  alan-1  64  4  0h31m  run  mcp
1631145  140  flynn  128  8  0h05m  run  TRON-JA307020

```

4. Run a simple job on the compute nodes.

At the conclusion of the installation process, the `CLEinstall` program provides suggestions for runtime commands and indicates how many compute nodes are available for use with the `aprun -n` option.

For `aprun` to work cleanly, the current working directory on the login node should also exist on the compute node. Change your current working directory to either `/tmp` or to a directory on a mounted Lustre file system.

For example, type the following.

```
crayadm@login:~> cd /tmp
crayadm@login:~> aprun -b -n 16 -N 1 /bin/cat /proc/sys/kernel/hostname
```

This command returns the hostname of each of the 16 compute nodes used to execute the program.

```
nid00010
nid00011
nid00012
nid00020
nid00016
nid00040
nid00052
nid00078
nid00084
nid00043
nid00046
nid00049
. . .
```

5. Test file system functionality. For example, if you have a Lustre file system named `/mylustmnt/filesystem`, type the following.

```
crayadm@login:~> cd /mylustmnt/filesystem
crayadm@login:/mylustremnt/filesystem> echo lustretest > testfile
crayadm@login:/mylustremnt/filesystem> aprun -b -n 5 -N 1 /bin/cat ./testfile
lustretest
lustretest
lustretest
lustretest
lustretest
Application 109 resources: utime ~0s, stime ~0s
```

6. Test the optional features that you have configured on your system.
 - a. To test RSIP functionality, log on to an RSIP client node (compute node) and ping the IP address of the SMW or other host external to the Cray system. For example, if `c0-0c0s7n2` is an RSIP client, type the following commands.

```
crayadm@login:~> exit
boot:~ # ssh root@c0-0c0s7n2
root@c0-0c0s7n2's password:
Welcome to the initramfs
# ping 172.30.14.55
172.30.14.55 is alive!
# exit
Connection to c0-0c0s7n2 closed.
boot:~ # exit
```

NOTE: RSIP clients on the compute nodes make connections to the RSIP server(s) during system boot. Initiation of these connections is staggered over a two minute window; during that time, connectivity over RSIP tunnels is unreliable. Avoid using RSIP services for three to four minutes following a system boot.

- b. To check the status of DVS, type the following command on the DVS server node.

```
crayadm@login:~> ssh root@nid00019 /etc/init.d/dvs status
DVS service: ..running
```

To test DVS functionality, invoke the `mount` command on any compute node.

```
crayadm@login:~> ssh root@c0-0c0s7n2 mount | grep dvs
/dvs-shared on /dvs type dvs
(rw,blksize=16384,nodename=c0-0c0s4n3,nocache,nodatasync,\
retry,userenv,clusterfs,maxnodes=1,nnodes=1)
```

Create a test file on the DVS mounted file system. For example, type the following.

```
crayadm@login:~> cd /dvs
crayadm@login:/dvs> echo dvstest > testfile
crayadm@login:/dvs> aprun -b -n 5 -N 1 /bin/cat ./testfile
dvstest
dvstest
dvstest
dvstest
dvstest
Application 121 resources: utime ~0s, stime ~0s
```

- Following a successful installation, the file `/etc/opt/cray/release/clerelease` is populated with the installed release level. For example,

```
crayadm@login:~> cat /etc/opt/cray/release/clerelease
5.2.UP04
```

If the preceding simple tests ran successfully, the system is operational. Cray recommends using the `xthotbackup` utility to create a backup of a newly updated or upgraded system. For more information, see the `xthotbackup(8)` man page.

Copy Boot Images and CLE Install Directory to the Passive SMW

- If boot images are stored as files, log on to `smw1` as `root` and copy the boot image to the other SMW. This manual copy operation speeds up future synchronization.

NOTE: In this command, replace `smw1` with the host name of the active SMW, and replace `smw2` with the host name of the passive SMW. Replace `bootimagedir` with the name of the boot image directory, and replace `file` with the name of the boot image.

```
smw1:~ # scp -p /bootimagedir/file smw2:/bootimagedir/file
```

IMPORTANT: The `bootimagedir` directory must already exist on the passive SMW.

- Copy the CLE install directory, `/home/crayadm/install.xtre1`, from the first SMW to a local directory on the second SMW (such as `/tmp`). Do not use `/home/crayadm` on the second SMW, because that would create local differences for this shared directory. Replace `xtre1` with the site-determined name specific to the release being installed.

Upgrade CLE Software on the Passive SMW

Use the following procedures to upgrade the CLE software on the passive SMW.

Install CLE Release Software on the SMW

Three DVDs are provided to install the CLE 5.2 release on a Cray system. The first is labeled `Cray CLE 5.2.UPnn Software` and contains software specific to Cray systems. Optionally, you may have an ISO image called `xc-sles11sp3-5.2.55d05.iso`, where `5.2.55` indicates the CLE release build level, and `d05` indicates the installer version.

To upgrade to the CLE 5.2 release from CLE or CLE 4.2 requires a second DVD labeled `Cray-CLEbase11sp3-yyyymmdd` and contains the CLE 5.2 base operating system, which is based on SLES 11 SP3. The third DVD is labeled `CentOS-6.5-x86_64-bin-DVD1.iso` and contains the CentOS 6.5 base operating system for CLE direct-attached Lustre (DAL) nodes.

Copy the Software to the SMW

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Mount the release media by using one of the following commands, depending on your media type.

If installing the release package from disk, place the `Cray CLE 5.2.UPnn Software` DVD in the CD/DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the release media using the ISO image, execute the following command, where `xc-sles11sp3-5.2.55d05.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro xc-sles11sp3-5.2.55d05.iso /media/cdrom
```

3. Copy all files to a directory on the SMW in `/home/crayadm/install.xtre1`, where `xtrel` is a site-determined name specific to the release being installed. For example:

```
smw:~# mkdir /home/crayadm/install.5.2.55
smw:~# cp -pr /media/cdrom/* /home/crayadm/install.5.2.55
```

4. Unmount the `Cray CLE 5.2.UPnn Software` media.

```
smw:~# umount /media/cdrom
```

5. For upgrading from CLE 5.1 or CLE 4.2 to CLE 5.2, you must mount the SLES 11 SP3 base media. Insert the `Cray-CLEbase11sp3` DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the base operating system media using the ISO image, execute the following command, where `Cray-CLEbase11SP3-yyyymmdd.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro Cray-CLEbase11SP3-yyyymmdd.iso /media/cdrom
```

6. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` image.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Install CLE on the SMW

1. As `root`, execute the `CRAYCLEinstall.sh` installation script to upgrade the Cray CLE software on the SMW.

```
smw:~# /home/crayadm/install.5.2.56/CRAYCLEinstall.sh \
-m /home/crayadm/install.5.2.56 -u -v -w
```

2. At the prompt `Do you wish to continue?`, type `y` and press `Enter`.

The output of the installation script is displayed to the console. If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

Finish the CLE Upgrade

After upgrading the CLE software on the passive SMW, use this procedure to finish the upgrade on the SMW HA system.

Add the following files to the synchronization list:

```
/var/opt/cray/install/shell_bootimage*
/var/opt/cray/install/networking_configuration-p*.json
```

For this procedure, see [Add Site-specific Files to the Synchronization List](#).

Configure PMDB Storage

Choose one of these options to configure shared storage for the Power Management Database (PMDB).

- [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172. Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.
- [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. Shared storage: A logical disk, configured as a LUN (Logical Unit) or logical volume on the boot RAID. The boot RAID must have sufficient space for `/var/lib/pgsql`.

Cray strongly recommends using either mirrored storage (preferred) or shared storage. An unshared PMDB is split across both SMWs; data collected before an SMW failover will be lost or not easily accessible after failover. For more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9.

Configure Mirrored Storage with DRBD for the PMDB

Prerequisites

IMPORTANT:

If mirrored storage becomes available after the PMDB has been configured for shared storage, use the procedure [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322 instead of this procedure.

Before beginning this procedure:

- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- Plan sufficient time for this procedure. Transferring the Power Management Database (PMDB) to a 1 TB disk requires about 10 hours. The SMW HA cluster should be in maintenance mode until the synchronization operation completes. The Cray system (compute and service nodes) can remain up and can run jobs during this period.
- Check `/etc/fstab` to ensure that there is no entry for `phy3`.
- If upgrading or updating the SMW HA system, ensure that the following RPMs are installed on both SMWs and that the version number is 8.4.4 or higher:

```
drbd-bash-completion-8.4.4-0.22.9
drbd-kmp-default-8.4.4_3.0.101_0.15-0.22.7
drbd-udev-8.4.4-0.22.9
drbd-utils-8.4.4-0.22.9
drbd-pacemaker-8.4.4-0.22.9
drbd-xen-8.4.4-0.22.9
drbd-8.4.4-0.22.9
```

If necessary, install or update any missing RPMs with `zypper install drbd`.

Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.

This procedure configures the network for DRBD, configures the DRBD disks, and transfers the PMDB data from local disk to the mirrored DRBD disks.

1. Add `eth5` to the network files.

- a. Log in as root on the first SMW (`smw1`).

```
workstation> ssh root@smw1
```

- b. On `smw1`, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.2/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- c. In a separate terminal session, log in as root on the other SMW (*smw2*).

```
workstation> ssh root@smw2
```

- d. On *smw2*, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.3/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

2. Reinitialize the eth5 interface on both SMWs.

```
smw1:~# ifdown eth5; sleep 1; ifup eth5
```

```
smw2:~# ifdown eth5; sleep 1; ifup eth5
```

3. Verify the IP addresses from *smw1*.

```
smw1:~# ping -c3 10.5.1.3
```

4. Configure the firewall to allow eth5 as an internal connection on both SMWs.

- Edit the file `/etc/sysconfig/SuSEfirewall12` on both *smw1* and *smw2*.
- Locate the line containing the `FW_DEV_INT` variable.
- If necessary, add `eth5` to the end of the `FW_DEV_INT` line.

```
FW_DEV_INT="eth1 eth2 eth3 eth4 eth5 lo"
```

- Save your changes and exit the editor on both SMWs.

5. Reinitialize the IP tables by executing the `/sbin/SuSEfirewall12` command on both SMWs.

```
smw1:~# /sbin/SuSEfirewall12
```

```
smw2:~# /sbin/SuSEfirewall12
```

6. On the active SMW only, add the new DRDB disk to the SMW HA configuration.

NOTE: The following examples assume that *smw1* is the active SMW.

- Verify that the device exists on both SMWs.
For Dell R-630 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

For Dell R815 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

- b. Determine if the dedicated disk for the PMDB must be formatted. In this procedure, this disk is referred to as PMDISK.

NOTE: If the PMDISK is already correctly formatted, skip to step 6.f on page 317.

This procedure assumes that a disk drive is available for use as a dedicated drive for the PMDB. The drive should be physically located within the rack-mount SMW at slot 4. The drive should be of the specification 1 TB 7.2K RPM SATA 3Gbps 2.5in HotPlug Hard Drive 342-1998, per the SMW Bill of Materials. On a Dell PowerEdge R815 the device for PMDISK

is /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0 On a Dell PowerEdge R630 the device for PMDISK is /dev/disk/by-path/pci-0000:03:00.0-scsi-0:4:0.

- c. Verify that the PMDISK is inserted into the SMW by entering the correct device name. This example is for a Dell R815.

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
255 heads, 63 sectors/track, 121601 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1
```

Device	Boot	Start	End	Blocks	Id	System
--------	------	-------	-----	--------	----	--------

- d. Create a new primary partition for the PMDISK, and write it to the partition table. If there are any existing partitions on this disk, manually delete them first.

```
smw:#fdisk \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)p
Partition number (1-4, default 1): 1
First sector (2048-1953525167, default 2048): [press return]
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-1953525167, default 1953525167): [press
return]
Using default value 1953525167
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.
```

- e. Verify that the partition has been created. This should be device /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
```

```

1000.2 GB, 1000204886016 bytes
81 heads, 63 sectors/track, 382818 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1

```

	Device	Boot	Start	End	Blocks	Id
System						
	/dev/disk/by-path/. . .-lun-0-part1		2048	1953525167	976761560	83
Linux						

- f. Navigate to the directory containing the SMWHAconfig command.

```
smw1:~# cd /opt/cray/ha-smw/default/hainst
```

- g. Execute SMWHAconfig to add the DRBD disk. For *disk-device*, specify the disk ID of the disk backing the DRBD disk, using either the by-name or by-path format for the device name. On a rack-mount SMW (either Dell R815 or R630), the DRBD disk is a partition on the disk in slot 4. On a Dell 815 this is /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1. On a Dell 630 it is /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1

```
smw1:~# ./SMWHAconfig --add_disk=pm-fs --device=/dev/drbd_r0 --directory=/var/lib/pgsql \
--pm_disk_name=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

7. Reboot the active SMW (*smw1*) and wait for it to boot completely.
8. Reboot the other SMW (*smw2*) and wait for it to boot completely.
9. Correct the permissions for the /var/lib/pgsql file on the active SMW.

```
smw1:~# chown postgres:postgres /var/lib/pgsql
smw1:~# chmod 750 /var/lib/pgsql
```

10. Put the SMW HA cluster into maintenance mode while waiting for the DRBD sync operation to complete. When *smw1* and *smw2* rejoin the cluster after rebooting, the primary DRBD disk (in *smw1*) synchronizes data to the secondary disk (in *smw2*). DRBD operates at the device level to synchronize the entire contents of the PMDB disk. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. The Cray system (service and compute nodes) can be booted and can run jobs during this period.

IMPORTANT:

Cray strongly recommends putting the SMW HA cluster into maintenance mode to prevent any failover during the sync operation. If a failover were to occur during this period, the newly-active SMW could have an incomplete copy of PMDB data.

- a. Put the SMW HA cluster into maintenance mode on *smw1*.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- b. Check the status of the DRBD sync operation with either `rcdrbd status` or `cat /proc/drbd`. The `rcdrbd` output is easier to read, but `/proc/drbd` contains more status information and includes an estimate of time to completion.

```

smw1:~# rcdbrd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res  cs          ro          ds          p
mounted          fstype
0:r0   SyncSource   Primary/Secondary  UpToDate/Inconsistent  C  /var/lib/
pgsql  ext3
...   sync'ed:      72.7%          (252512/922140)M

```

```

smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2
ua:0 ap:0 ep:1 wo:f oos:260636656
[=====>.....] sync'ed: 72.4% (254524/922140)M
finish: 2:21:07 speed: 30,768 (29,720) K/sec

```

For an explanation of the status information in `/proc/drbd`, see the DRDB User's Guide at linbit.com: <http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd>.

- When the DRBD sync operation finishes, bring the HA cluster out of maintenance mode on `smw1`.

```

smw1:~# crm configure property maintenance-mode=false 2> /dev/null

```

- Examine the output of `crm status` to ensure that the `ip_drbd_pgsql` is started on `smw1` and that the Masters and Slaves entries for `ms_drbd_pgsql` display the SMW host names (`smw1` and `smw2`).

```

smw1:~# crm status
Last updated: Thu Jan 22 18:40:21 2015
Last change: Thu Jan 22 11:51:36 2015 by hacluster via crmd on smw1
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.11-3ca8c3b
2 Nodes configured, 2 expected votes
23 Resources configured

Online: [ smw1 smw2 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
.
.
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):              Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql   (lsb:postgresql):              Started smw1
  mysqld       (ocf::heartbeat:mysql):         Started smw1
  ip_drbd_pgsql (ocf::heartbeat:IPaddr2):       Started smw1
Master/Slave Set: ms_drbd_pgsql [drbd_pgsql]

```

```
Masters: [ smw1 ]
Slaves: [ smw2 ]
```

Configure Shared Storage on the Boot RAID for the PMDB

Prerequisites

The SMW HA system can be configured to store the Power Management Database (PMDB) on shared storage, a logical disk configured as a LUN (Logical Unit) or logical volume on the boot RAID.

IMPORTANT: Cray strongly recommends using mirrored storage, if available, for the PMDB; for more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9. To move the PMDB from shared storage to mirrored storage, see [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322.

Before beginning this procedure:

- Ensure that the boot RAID contains a LUN for the PMDB with sufficient space for the data. Use the following command to check the size of `/var/lib/pgsql` on the local disk:

```
smw1:~ # du -hs /var/lib/pgsql
```

- Check that the boot RAID is connected.
- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- To capture typescript output from this procedure, do not use a typescript session running directly on the SMW. To save the output of this procedure, use the `script` command to start the typescript session on your local workstation before logging into the SMW, as in this example:

```
workstation> script -af my_output_file
Script started, file is my_output_file
workstation> ssh crayadm@smw1
```

Use this procedure to configure the RAID disk and transfer the power management data base (PMDB) to the power management disk on the shared boot RAID.

- Shut down the Cray system by typing the following command as `crayadm` on the active SMW (`smw1`).

```
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
```

- Log into the active SMW as `root`, either at the console or by using the actual (not virtual) host name.

IMPORTANT: You must log in directly as `root`. Do not use `su` from a different SMW account such as `crayadm`.

- Change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

- Use the `SMWHAconfig` command to move the PMDB and configure the required HA resources. In the following command, replace `scsi-xxxxxxxx` with the persistent device name for the PMDB directory on the boot RAID.

```
smw1:~ # ./SMWHAconfig --add_disk=pm-fs \
--device=/dev/disk/by-id/scsi-xxxxxxxx --directory=/var/lib/pgsql
```

This command mounts the PMDB directory (`/var/lib/pgsql`) to the boot RAID, copies the PMDB data, and configures the HA resources `pm-fs` and `postgresqld`.

5. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a ping command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

6. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a ping command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

7. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons   (lsb:rsms):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog (lsb:cray-syslog):             Started smw1
  homedir    (ocf::heartbeat:Filesystem):    Started smw1
  md-fs     (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs     (ocf::heartbeat:Filesystem):    Started smw1
  postgresqld (lsb:postgresql):             Started smw1
  mysqld    (ocf::heartbeat:mysql):        Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.

- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

8. Verify that the Power Management Database is on the boot RAID and that the required PMDB resources are running.
- a. Examine the log file `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out` to verify that the Power Management Database disk appears in the `Cluster RAID Disks` section (at the end of the file), as in this example.

```
----- Cluster RAID Disks -----
07-07 20:47 INFO   MYSQL Database disk = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   Log disk            = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   /home disk         = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   PM database disk   = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO   ***** Ending of HA software add_disk *****
```

- b. Ensure that the power management file system is mounted by checking for `/var/lib/pgsql` in the output of the `df` command.

```
smw1:~ # df
Filesystem      1K-blocks      Used Available Use% Mounted on
/dev/sda2       120811676    82225412  32449332  72% /
udev            16433608         756  16432852   1% /dev
tmpfs           16433608     37560  16396048   1% /dev/shm
/dev/sdo        483807768  197536596  261695172  44% /var/opt/cray/disk/1
/dev/sdp        100791728   66682228  28989500  70% /home
/dev/sdq        100791728   484632   95187096   1% /var/lib/mysql
/dev/sdr        30237648    692540  28009108   3% /var/lib/pgsql
```

- c. Check the output of `crm_mon` to ensure that the `pm-fs` and `postgresqd` resources are running.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
```

```

Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.

```

```
Online: [ smw1 smw2 ]
```

```

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                    Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                    Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir      (ocf::heartbeat:Filesystem):    Started smw1
  md-fs        (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):    Started smw1
  postgresql   (lsb:postgresql):             Started smw1
  mysqld       (ocf::heartbeat:mysql):        Started smw1

```

Migrate PMDB Data from the Boot RAID to Mirrored Storage

Prerequisites

Before beginning this procedure:

- Ensure that the mirrored PMDB disk has been configured as specified in [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172.
- Identify the device name of the boot RAID partition containin the Power Management Database (PDMB).

Use the following procedure to move the PMDB data from shared storage on the boot RAID to the mirrored storage on the DRBD disk.

1. Log into the active SMW as `root`.
2. Put the cluster in maintenance mode.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

3. Stop `rsms`.

```
smw1:~# rsms stop
```

4. Stop `postgresql`.

```
smw1:~# /etc/init.d/postgresql stop
```

5. Mount the boot RAID partition previously used by the PMDB.

```
smw1:~# mount boot_RAID_partition /mnt/pgsql_tmp
```

6. Back up the existing copy of /var/lib/pgsql, if possible.

```
smw1:~# cp -pr /var/lib/pgsql /var/lib/pgsql-backup
```

7. Remove the existing contents of /var/lib/pgsql on the mirrored disk.

```
smw1:~# rm -rf /var/lib/pgsql/*
```

8. Copy the PMDB contents from the boot RAID partition to /var/lib/pgsql.

```
smw1:~# cp -pr /mnt/pgsql_tmp/* /var/lib/pgsql
```

9. Start postgresql.

```
smw1:~# /etc/init.d/postgresql start
```

10. Check the postgresql status.

```
smw1:~# /etc/init.d/postgresql status
Checking for PostgreSQL
9.1.12: running
```

11. Start rsms.

```
smw1:~# rsms start
```

12. Inspect the status of the rsms daemons and the contents of /var/opt/cray/log/power_management-YYYYMMDD, where YYYYMMDD is today's date. If xtpmd is running and no database errors are noted, the transfer went properly.

```
smw1:~# rsms status
cluster is in maintenance mode and daemons are not under cluster control
Checking for RSMS service:
erd.. running
Checking for RSMS service:
erdh.. running
Checking for RSMS service:
sm.. running
Checking for RSMS service:
nm.. running
Checking for RSMS service:
bm.. running
Checking for RSMS service:
sedc_manager.. running
Checking for RSMS service:
cm.. running
Checking for RSMS service:
xtpmd.. running
Checking for RSMS service:
erfsd.. running
```

```
Checking for RSMS service:
xtremoted..                                running
```

13. If the `rsms` status is good, remove the backup of `/var/lib/pgsql`.
14. Wait for the PMDB to sync completely. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. Check the status of the DRBD sync operation with either `rcdrbd status` or `cat /proc/drbd`. The `rcdrbd` output is easier to read, but `/proc/drbd` contains more status information and includes an estimate of time to completion.

```
smw1:~# rcdrbd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res  cs          ro          ds          p  mounted
fstype
0:r0   SyncSource   Primary/Secondary  UpToDate/Inconsistent  C  /var/lib/pgsql
ext3
...   sync'ed:      72.7%          (252512/922140)M
```

```
smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
 0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
   ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2 ua:
0 ap:0 ep:1 wo:f oos:260636656
   [=====>.....] sync'ed: 72.4% (254524/922140)M
   finish: 2:21:07 speed: 30,768 (29,720) K/sec
```

15. Take the cluster out of maintenance mode.

```
smw1:~# crm configure property maintenance-mode=false 2 > /dev/null
```

Update the Cray SMW HA System

Cray provides periodic updates to the SMW, CLE, and SMW HA software releases. A software update on an SMW HA system involves installing the next minor release for all necessary software components (the SMW, CLE, and SMW HA software). In an update release, the minor version number changes; for example, from SMW SLEHA 11 SP3 to SMW SLEHA 11 SP3 UP01.

The update procedures do not change the base operating system version running on your SMW.

Note the requirements for updating an SMW HA system:

- Both SMWs must be running the SUSE Linux Enterprise Server version 11 Service Pack 3 (SLES 11 SP3) SMW base operating system and a release of SMW 7.2 or later.
- Update the SMW software and SMW HA software in the same upgrade session. Cray recommends updating in the following order:
 1. SMW software
 2. SMW HA software
- For each update release package, update the active SMW first, then update the passive SMW. Do not update both SMWs at the same time.
- The CLE software can be updated either before or after the SMW and SMW HA updates.

The update procedures use the following conventions to refer to the SMWs:

- The host name `smw1` specifies the currently active SMW. In examples, the prompt `smw1:~ #` shows a command that runs on this SMW.
- The host name `smw2` specifies the currently passive SMW. In examples, the prompt `smw2:~ #` shows a command that runs on this SMW.
- The host name `virtual-smw` host name specifies the active SMW (which could be either `smw1` or `smw2`). This virtual host name was defined during initial installation.

Before You Start an SMW HA Update

To prepare to update the software on an SMW HA system, do the following:

- Read the *SMW HA Release Notes* and the *SMW HA README* to confirm the required versions for the operating system, SMW, and SMW HA software release, as well as the supported upgrade or update paths. These documents are provided with the SMW HA release package.
- Read the *SMW HA README* and *SMW HA Errata* to determine whether there are any corrections to the upgrade or update procedures. These documents are included in the SMW SLEHA upgrade or update directory.
- Read any Field Notices (FNs) related to kernel security fixes.

IMPORTANT: Kernel 3.0.101-0.461 (provided in FN-6029) or later is required for an SMW HA system. There is a SLES kernel dependency on the `ocfs2-kmp-default` RPM package that will prevent some SLES HA RPMs from being installed unless this kernel update has been applied.

- Back up the current SMW and SMW HA software before installing the upgrade or update packages. For more information, see [R815 SMW: Create an SMW Bootable Backup Drive](#) on page 182.
- Cray recommends checking all file systems with `fsck` before beginning an upgrade or update, because an SMW HA system requires several reboots during this procedure. Exceeding a file-system mount count would delay a reboot by triggering an automatic file-system check.
- Identify any local changes to the list of synchronized files and directories in `/etc/csync2/csync2_cray.cfg`. The installation procedure saves local changes in a temporary file. You will restore those changes in a post-installation step.
- Plan sufficient time. An SMW HA system requires more time to upgrade or update, as compared to a system with a single SMW, because you must install the software on both SMWs. Allow at least two hours of additional time.

Update SMW Software

For a system running the SMW 7.2.UP00 (or later) software, the procedures in this section are required for updates to the SMW 7.2 release package. The following procedures are required:

1. [Prepare the SMW HA System for an SMW Update](#) on page 326
2. [Update SMW software on the Active SMW](#) on page 328 (including steps to restart the cluster resources)
3. [Update SMW Software on the Passive SMW](#) on page 344
4. [Finish the SMW Update](#) on page 355

When updating the SMW software, update the active SMW first, then update the passive SMW. Do not update both SMWs at the same time.

NOTE:

If the system is running the SMW 7.1.UP01 (or earlier) software, upgrade the software using the procedures in [Upgrade the SMW Software](#) on page 243.

Prepare the SMW HA System for an SMW Update

1. Log on to each SMW as `root`.
2. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, shut down the VNC server.

```
smw1:~ # /etc/init.d/vnc stop
```

3. Determine whether the `postgresql` service is currently on or off, and record this state. After completing the upgrade or update, you will return the `postgresql` service to the same state.

```
smw1:~ # chkconfig postgresql
postgresql state
```

- Find the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -1 | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

NOTE: The examples in this procedure assume that `smw1` is currently the active SMW.

- Record the iDRAC IP address of both SMWs in case you need to power-cycle either SMW.

Usually, the iDRAC host name follows the naming convention `hostname-drac`. For example, if the host names are `smw1` and `smw2`, the iDRAC host names would be `smw1-drac` and `smw2-drac`. Use the following ping commands to display the iDRAC IP addresses.

NOTE: In these commands, replace `smw1-drac` with the host name of the iDRAC on the active SMW. Replace `smw2-drac` with the host name of the iDRAC on the passive SMW.

```
smw1:~ # ping smw1-drac
PING smw1-drac.us.cray.com (172.31.73.77) 56(84) bytes of data.
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=1 ttl=64
time=1.85 ms
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=2 ttl=64
time=0.398 ms
64 bytes from smw1-drac.us.cray.com (172.31.73.77): icmp_seq=3 ttl=64
time=0.408 ms
...

smw1:~ # ping smw2-drac
PING smw2-drac.us.cray.com (172.31.73.79) 56(84) bytes of data.
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=1 ttl=64
time=1.85 ms
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=2 ttl=64
time=0.398 ms
64 bytes from smw2-drac.us.cray.com (172.31.73.79): icmp_seq=3 ttl=64
time=0.408 ms
...
```

- Shut down the Cray system as `crayadm` on the active SMW.

```
smw1:~ # su - crayadm smw1
...
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
...
crayadm@smw1:~> exit
smw1:~ #
```

- As `root` on the active SMW, stop file synchronizing.

```
smw1:~ # crm resource stop fsync
```

- On the active SMW, turn on maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=true 2> /dev/null
```

- Determine the persistent device name of the `/home` directory.

```
smw1:~ # crm configure show | grep device | awk '{print $2 " " $3}' | \
sed 's/"//g'
```

```
device=/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9c01 directory=/home
device=/dev/disk/by-id/scsi-360080e500023bff6000006b3515d9bdf directory=/var/
lib/mysql
device=/dev/disk/by-id/scsi-360080e500023bff6000006b1515d9bc9 directory=/var/
opt/cray/disk/1
```

Update SMW software on the Active SMW

Use the following procedures to to update the SMW software on the active SMW.

Update or Upgrade the Cray SMW Software

Cray provides periodic updates to each System Management Workstation (SMW) release, as well as upgrade releases. In an upgrade release, the major and/or minor version number will change, for example from 7.1.UP01 to 7.2.UP00. In an update release, only the minor version (numbers following *UP*) will change.

Follow the procedures in this chapter to install an `SMW 7.2UP04` package. The procedures provided in this chapter do not change the base operating system version running on the SMW.



CAUTION: The SMW must be running the SUSE Linux Enterprise Server version 11 Service Pack 3 (SLES 11 SP3) SMW base operating system and a release of SMW 7.1 or later in order to perform the following update/upgrade procedures.

Prepare to Upgrade or Update SMW Software

IMPORTANT: For a system configured for SMW high availability (HA) with the SMW failover feature, prepare both SMWs for an upgrade.

- Determine which SLES version is running on the SMW by executing the following command:

```
crayadm@smw> cat /etc/SuSE-release
```

- Read the *SMW README* and *SMW Errata* provided in the SMW update directory for any changes to upgrade or update procedures.
- Read the Field Notices (FN) related to kernel security fixes, and apply any needed changes before continuing with the installation.
- If local changes have been made to the file `/opt/cray/hss/default/etc/sedc_srv.ini`, note the following information.
 - Cray software manages this file as a symbolic link to `/opt/cray/hss/default/etc/sedc_srv.ini.xtek` (Cray XE and Cray XK systems only) or to `/opt/cray/hss/default/etc/sedc_srv.ini.cascade` (Cray XC30 systems only). The following actions are taken during software updates:
 - If the symbolic link exists, it is not altered.
 - If the symbolic link does not exist, it is created as specified above.
 - If `sedc_srv.ini` exists not as a symbolic link but as a regular file, it is renamed to `/opt/cray/hss/default/etc/sedc_srv.ini-YYYYMMDDhhmmss` (where `YYYYMMDDhhmmss` is the date and time the file was renamed) and a new symbolic link is created.
 - Before beginning the upgrade, copy `sedc_srv.ini` to a new site-specific file and change the symbolic link to point to that file. After the software update, compare the local file to the distribution `sedc_srv.ini.xtek` or `sedc_srv.ini.cascade` file for any changes that should be merged into the locally modified file.

- If local changes have been made to any automation files, such as `/opt/cray/hss/default/etc/auto.xtshutdown`, back them up before beginning the SMW upgrade.
- For Cray XC Systems: To retain any power management profiles that were created, back up all of the files in the `/opt/cray/hss/7.1.0/pm/profiles` directory before beginning the SMW upgrade.
- For liquid-cooled Cray XC30 Systems only: One or more patches to the SMW 7.1.UP01 release may have installed an `hss.ini` file on the system. This file, and a file containing the default values, `hss.ini.dist` are located in `/opt/tftpboot/ccrd`.
 - If `hss.ini` does not have local changes, delete this file before beginning the update installation, and an active `hss.ini` file will be created as part of the installation.
 - If `hss.ini` has local changes, save a local copy of `hss.ini` to another location. After installation is complete, copy `hss.ini.dist` to `hss.ini` and re-create the local changes in the new `hss.ini` file.
- If using the Cray simple event correlator (SEC) and the `/opt/cray/default/SEC_VARIABLES` file has local changes, make a backup copy of this file before beginning the upgrade or update. For more information, see *Configure Cray SEC Software (S-2542)*.
- If `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` has local changes, the local changes are saved during the upgrade procedure. Also, if `/opt/cray/hss/default/etc/xtnlrd.ini` has local changes and the new release includes an updated `xtnlrd.ini` file, the local version of the file is preserved during the upgrade and the new file is installed as `xtnlrd.ini.rpmnew`. After the upgrade, compare the two files and merge any changes into the local `xtnlrd.ini` file.
- Update the `properties.local` file. (See [Update the properties.local File](#))
- Back up the current software. (See [Back Up the Current Software](#) on page 24).

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `phy7` and is slot 0, and the bootable backup disk is `phy6` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 332; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
  /dev/sda
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x443322110700000-lun-0
Physical slot 1:
  /dev/sdc
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RD7
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x443322110600000-lun-0
Physical slot 2:
  /dev/sdd
  /dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
  /dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x443322110500000-lun-0
Physical slot 3:
  /dev/sdb
  /dev/disk/by-id/ata-ST9500620NS_9XF0665V
  /dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
  /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x443322110400000-lun-0
Physical slot 4:
  NOT INSTALLED
Physical slot 5:
  NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id` device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive

into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and grub) drive names, the `device.map` mapping file used by grub should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 330 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 330, the `fstab` lines would change from:

```
/dev/sda1 swap swap defaults 0 0
/dev/sda2 / ext3 acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 swap swap defaults 0 0
```

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 /      ext3  acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the grub bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the grub utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the grub utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported.  For the first word, TAB
  lists possible command completions.  Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
  Checking if "/boot/grub/stage1" exists... yes
  Checking if "/boot/grub/stage2" exists... yes
  Checking if "/boot/grub/e2fs_stage1_5" exists... yes
  Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
  Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
  Boot  Start      End    Blocks  Id System
        63  16771859   8385898+  82  Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
  Boot  Start      End    Blocks  Id System
        * 16771860 312576704 147902422+  83  Linux
```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSs
    (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080

           Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
           Boot   Start       End     Blocks  Id  System
                63  16771859     83828   82    Linux

swap / Solaris
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
           Boot   Start       End     Blocks  Id  System
                167719 312581807 156207044+ 83    Linux

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807

Command (m for help): w
```

The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

- c. Display the boot backup disk partition layout and confirm it matches the phy7 sector information.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
```

4. Initialize the swap device.

```
smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
(DOS partition table detected). Use -f to force.
Setting up swapspace version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed
```

5. Create a new file system on the backup drive root partition by executing the mkfs command.

```
smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 37 mounts or
180 days, whichever comes first. Use tune2fs -c or -i to override.
```

6. Mount the new backup root file system on /mnt.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem          1K-blocks      Used Available Use% Mounted on
/dev/sda2            303528624    6438700 281671544   3% /
udev                 1030332         116  1030216   1% /dev
/dev/sdb2            306128812    195568 290505224   1% /mnt
```

The running root file system device is the one mounted on /.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's fstab and menu.lst files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the root and swap partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.

For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the
  possible
  completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
grub> root (hd2,1)
root (hd2,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
embedded.
succeeded
Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
Done.
grub> quit
```

IMPORTANT: For R815 SMWs, `grub` recreates `device.map` with the short names, not the persistent names. Do not trust the `/dev/sdx` names. Always use `find` when executing `grub` because it is possible that `grub root` may not be `hd2` the next time `grub` is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

Shut Down the Cray System

1. Log on to the SMW as `crayadm` and confirm the Cray system is shut down.

```
crayadm@smw:~> ping boot
```

If the command responds with "alive", then it is up and needs to be shut down.

2. Shut down the system by typing the following command.

```
crayadm@smw> xtbootsys -s last -a auto.xtshutdown
```

For more information about using automation files, see the `xtbootsys(8)` man page.

Update the SMW Software and Configuration

1. Open a terminal window, and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. Mount the release media by using one of the following commands, depending on the media type.

- To install the update package from disk, place the SMW 7.2UP04 Software DVD in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as `/tmp/SMW_version` on the SMW and then substitute this path for `/media/cdrom` in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the `smw-image` ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

3. If `postfix` is not configured on the SMW, skip this step:

To prevent the `master.cf` and `main.cf` `postfix` configuration files from being recreated during software updates or fixes, ensure the following setting in the `/etc/sysconfig/mail` file on the SMW is set to "no":

```
MAIL_CREATE_CONFIG="no"
```

4. To see what SMW software will be updated in this new release, execute these commands prior to doing the update. This information is gathered, displayed, and contained in the log files during the `SMWinstall` process.
 - a. Check security and recommended updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GS
```

- b. Check Cray software updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GV
```

5. Create a new copy of the `SMWinstall.conf` configuration file and modify the new copy of the `SMWinstall.conf` file with site-specific requirements. Only `root` can modify the `SMWinstall.conf` configuration file. The `SMWinstall.conf` configuration file is created during the installation process by copying the `SMWinstall.conf` template from the distribution media. By default, the SMW configuration file is placed in `/home/crayadm/SMWinstall.conf`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
smw# chmod 644 /home/crayadm/SMWinstall.conf
smw# vi /home/crayadm/SMWinstall.conf
```

IMPORTANT: For an SMW HA system, define this variable as if the system was a standalone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.

For a description of the contents of the `SMWinstall.conf` file, see the `SMWinstall.conf(5)` man page.

6. Update the software with `SMWinstall`. `SMWinstall` checks for any inconsistency between the system and the `SMWinstall.conf` file settings, prompts for the root MySQL database password, and stores its log files in `/var/adm/cray/logs`.

```
smw# /media/cdrom/SMWinstall
...
```

```
Please enter your root DB password:
Please confirm your root DB password:
Password confirmed.
```

When `SMWinstall` finishes, it will suggest a reboot of the SMW.

7. If necessary, restore the locally modified versions of the following files.
 - a. If the installation site had locally modified versions of `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` before this SMW update, restore the local modifications to these files. During the upgrade procedure, the old files are saved in `/etc/syslog-ng/syslog-ng.conf-YYYYMMDDhhmm` and `/etc/rsyslog.conf.rpmsave`.
 - b. If the installation site had local modifications to `/opt/cray/hss/default/etc/auto.xtshutdown` before this SMW update, restore the local modifications to this file.
 - c. If the installation site had power management profiles, copy the files that were backed up from the `/opt/cray/hss/7.1.0/pm/profiles` directory into the new `/opt/cray/hss/default/pm/profiles` directory. See *Monitoring and Managing Power Consumption on the Cray XC30 System (S-0043)* for instructions on validating the restored profiles.
 - d. If the installation site had local modifications to `/opt/cray/hss/default/etc/sedc_srv.ini` before this SMW update, locate the destination of this symbolic link (see [Prepare to Upgrade or Update SMW Software](#) on page 245), compare the content of the local file to the distributed version of the file, and update the local file appropriately.
8. If the installation site has an `/opt/cray/hss/default/etc/xtdiscover.ini` file, the SMW update process does not overwrite an existing `xtdiscover.ini` file; the new version is named `xtdiscover.ini.dist`. Compare the content of the new `xtdiscover.ini.dist` with the original `xtdiscover.ini` file, and update the `xtdiscover.ini` file appropriately.

NOTE: If the `xtdiscover.ini` file does not exist, then the `xtdiscover.ini.dist` file is copied to the `xtdiscover.ini` file.

9. Unmount the SMW 7.2UP04 Software media.

```
smw# umount /media/cdrom
```

10. If using the update disk, eject the SMW 7.2UP04 Software DVD.

```
smw# eject
```

11. Reboot the SMW.

```
smw# reboot
```

For a Cray XE or Cray XK system, continue to [Update the L0 and L1 Firmware](#).

Start Cluster Resources

Prerequisites

After updating the SMW software and configuration and rebooting the SMW, use this procedure to start the cluster resources.

NOTE: The examples in this procedure assume that `smw1` was the active SMW at the start of the update.

1. Wait for the SMW to finish rebooting at the end of the previous procedure.
2. Log into the active SMW as `root`.
3. Disable maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=false 2> /dev/null
```

4. Wait for the operation to complete.

```
smw1:~ # sleep 30
```

5. Enable maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=true 2> /dev/null
```

For Cray XC Series Systems Only: Update the BC and CC Firmware

Cray XC Series images for the cabinet controller (CC) and blade controller (BC) are always downloaded over the HSS network. Any updated firmware will be placed in `/opt/cray/hss-images/...` as part of the installation or update process. In order to boot the updated firmware, the `hss_make_default_initrd` script must be run, all CCs rebooted, and all BCs rebooted and power cycled.

IMPORTANT: If an installation step fails because of a hardware issue, such as a cabinet failing to power up, when that issue is resolved, go back to the last successful step in the installation procedure and continue from there. Do not skip steps or continue out of order.

1. Update the controller boot image.

The version used in the command argument for `hss_make_default_initrd` should match that of the version specified in the `lsb-cray-hss` line in the output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Mon Sep 14 00:57:20 CDT 2015 on hssbld0 by bwdev
lsb-cray-hss-7.2.0-1.0702.37336.662
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.37336.662
::: Verifying base RPM list to the manifest
::: Installing filesystem hierarchy
::: Installing busybox
::: Installing base RPMs
::: Installing ssh
::: Removing /etc/securetty installed by the pam RPM
::: Installing sshfs
::: Installing rsh
::: Modifying /etc/pam.d/rlogin to remove securetty checking
::: Modifying /etc/pam.d/rsh to remove rhosts and nologin checking (Bug #779466)
::: Installing rsync
::: Installing atftp
::: Installing tcpdump
::: Installing ethtool
::: Installing syslog-ng
::: Installing logrotate
::: Installing ntp/ntpd
::: Installing strace
::: Installing screen
::: Installing minicom
::: Installing ppp
::: Installing mtd-utils
::: Installing /init
::: Installing file.rpm
::: Installing libgmodule
::: Installing Midnight Commander
::: Installing cray-viper
::: Installing spread
::: Installing coreboot-utils
::: Clearing init.d to be replaced by cray-hss32-filesystem
::: Creating initial etc files needed for root creation
::: Installing Cray kernel
::: Installing latest Cray kernel modules
::: Clearing select /boot items
::: Installing boot-parameters
::: Installing cray-hss32-scripts
::: Installing lsb-cray-hss-controllers
::: Installing cray-libconfig
::: Installing cray-bdm
::: Installing cray-play_xsvf
::: Removing unwanted files from the root

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.37336.662.

Running hssclone.
Image Clone Complete: /opt/cray/hss-images/image-7.2.0-1.0702.37336.662
Running hsspackage.
copying image
copying modules
```

```

running depmod
creating load file: /opt/cray/hss-images/default/HSS32.load
compressing initrd.img
Creating pxelinux.0 symlink
Running hssbootlink.
linking /opt/cray/hss-images/default/HSS32/bzImage-3.0.76-0.11.1_1.0702.8867-
cray_hss32 /opt/tftpboot/bzImage
linking /opt/cray/hss-images/default/HSS32/parameters /opt/tftpboot/
pxelinux.cfg/default
linking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img

```

2. Power down the system.

```
smw# xtcli power down s0
```

3. Reboot the cabinet controllers, then ensure that all cabinet controllers are up.

```

smw# xtccreboot -c all
xtccreboot: reboot sent to specified CCs
smw# xtalive -l cc

```

4. Power up the system.

```
smw# xtcli power up s0
```

Note that at this point the `xtcli` status output shows that all nodes are "off", because they have not yet been bounced.

5. Run the `xtdiscover` command to ensure that any changes made to the HSS database schema for new features are captured.

```
smw# xtdiscover
```

6. Exit from the `root` login.

```
smw# exit
```

7. Run the `rtr --discover` command to determine the exact configuration of the HSN.

```
crayadm@smw> rtr --discover
```

If the system was not bounced previously, the following message may be displayed:

```
System was not bounced in diagnostic mode, should I re-bounce? Continue (y/n)?
```

If so, respond with `y`.

8. Update the firmware. Execute the `xtzap` command to update the components.



CAUTION: The `xtzap` command is normally intended for use by Cray Service personnel only. Improper use of this restricted command can cause serious damage to the computer system.

```
crayadm@smw> xtzap -r -v s0
```

IMPORTANT: The Cray XC30 system also requires an update to the NVIDIA® BIOS (nvBIOS) for the NVIDIA K20X graphics processing units (GPUs). This update is done after CLE has been booted. For more information, see *CLE Installation and Configuration Guide (S-2444)*.

9. Use the output of the `xtzap` command to determine if any components need to be flashed.

While the `xtzap -a` command can be used to update all components with a single command, it may be faster to use the `xtzap -blade` command when only blade types need to be updated, or the `xtzap -t` command when only a single type needs to be updated. On larger systems, this can be a significant time savings.

This is the list of all cabinet level components:

```
cc_mc (CC Microcontroller)
cc_bios (CC Tolapai BIOS)
cc_fpga (CC FPGA)
chia_fpga (CHIA FPGA)
```

This is a list of all blade level components:

```
cbb_mc (CBB BC Microcontroller)
ibb_mc (IBB BC Microcontroller)
anc_mc (ANC BC Microcontroller)
bc_bios (BC Tolapai BIOS)
lod_fpga (LOD FPGA)
node_bios (Node BIOS)
loc_fpga (LOC FPGA)
qloc_fpga (QLOC FPGA)
```

If the output of the `xtzap` command shows that only a specific type needs to be updated, then use the `-t` option with that type (this example uses the `node_bios` type).

```
crayadm@smw> xtzap -t node_bios s0
```

If the output of the `xtzap` command shows that only blade component types need to be updated, then use the `-b` option:

```
crayadm@smw> xtzap -b s0
```

If the output of the `xtzap` command shows that both blade- and cabinet-level component types need to be updated, or if there is uncertainty about what needs to be updated, then use the `-a` option:

```
crayadm@smw> xtzap -a s0
```

10. Execute the `xtzap -r -v s0` command again; all firmware revisions should report correctly, except `node_bios`; `node_bios` will display as "NOT_FOUND" until after the `xtbounce --linktune` command is run.

```
crayadm@smw> xtzap -r -v s0
```

11. Execute the `xtbounce --linktune` command, which forces `xtbounce` to do full tuning on the system.

```
crayadm@smw> xtbounce --linktune=all s0
```

Continue with [Confirm the SMW is Communicating with System Hardware](#) on page 43.

Update SMW Software on the Boot Root and Shared Root

This procedure uses the `SMWinstallCLE` script to update the SMW software on the boot root and shared root for systems already running the Cray Linux Environment (CLE) software. The RPMs that `SMWinstallCLE` installs on the boot root and shared root will also be installed when `CLEinstall` runs during a CLE update.

TIP: Use this procedure only if the plan is to boot CLE after the SMW update but before updating the CLE software. Otherwise, if the plan is to update CLE software without booting CLE after the SMW update, it is safe to skip this procedure.

For more information about the `SMWinstallCLE` script, see the `SMWinstallCLE(8)` man page.

- As `root`, mount the release media by using one of the following commands, depending on the media type.
 - To install the update package from disk, place the `SMW 7.2UP04 Software DVD` in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as `/tmp/SMW_version` on the SMW and then substitute this path for `/media/cdrom` in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the `smw-image` ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

- Update the `label_name` system set from the `/etc/sysset.conf` system set configuration file. In the following steps it is assumed that the label `label_name` is described in the `/etc/sysset.conf` system set configuration file. See the `sysset.conf(5)` man page for additional information about the `/etc/sysset.conf` file. For more detailed information about `SMWinstallCLE`, see the `SMWinstallCLE(8)` man page.

```
smw# /media/cdrom/utils/SMWinstallCLE --label=label_name
```

NOTE: The `SMWinstallCLE` command checks whether the boot node is booted. If it is booted, `SMWinstallCLE` prompts for confirmation that the system set being changed is not the one booted.

```
HH:MM:SS WARNING: Your bootnode is booted. Please confirm that the system
set you
intend to update is not booted.
Do you wish to proceed? [n] y
```

- Unmount the `SMW 7.2UP04 Software` media.

```
smw# umount /media/cdrom
```

- If using the update disk, eject the `SMW 7.2UP04 Software DVD`.

```
smw# eject
```

The SMW software is now updated. If the firewall is not yet configured, see [Set Up the SUSE Firewall and IP Tables](#) on page 45. Then continue to install the CLE software using [CLE Installation and Configuration Guide \(S-2444\)](#), which is provided with the CLE release package.

NOTE: To reconfigure a `LOGDISK`, `PMDISK`, or `DBDISK` when it is necessary to replace a failed drive with a new drive on a rack-mount SMW, see [Replace a Failed Disk Drive](#).

Configure the Simple Event Correlator (SEC)

The System Management Workstation (SMW) 7.2.UP04 release includes the Open Source simple event correlator (SEC) package, `sec-2.7.0`, and an SEC support package, `cray-sec-version`. The SEC support package contains control scripts to manage the starting and stopping of SEC around a Cray mainframe boot session, in addition to other utilities.

To use the Cray SEC, see *Configure SEC Software (S-2542)* for configuration procedures.

Update SMW Software on the Passive SMW

Before updating the SMW software on the passive SMW, use this procedure to prepare for the update.

The examples in this procedure assume that `smw1` was the active SMW at the start of the update.

1. Log into each SMW as `root`.

2. On the active SMW, turn off maintenance mode.

```
smw1:~ # crm configure property maintenance-mode=false 2> /dev/null
```

3. On the active SMW, force a failover to the passive SMW, then wait 30 seconds for the failover operation to complete.

```
smw1:~ # crm node standby
smw1:~ # sleep 30
```

4. On the other SMW (which is now the active one), clear the resource failcounts.

```
smw2:~ # clear_failcounts
```

5. Bring the first SMW online.

NOTE: Replace `smw1` with the host name of the SMW that was active at the start of this procedure.

```
smw2:~ # crm node online smw1
```

6. On the second SMW (`smw2`), turn on maintenance mode.

```
smw2:~ # crm configure property maintenance-mode=true 2> /dev/null
```

R815 SMW: Create an SMW Bootable Backup Drive

This procedure creates a bootable backup drive for a Dell R815 SMW in order to replace the primary drive if the primary drive fails. When these steps are completed, the backup drive on the SMW will be a bootable replacement for the primary drive when the backup drive is plugged in as or cabled as the primary drive.

IMPORTANT: The disk device names shown in this procedure are only examples. Substitute the actual disk device names for the actual system. The boot disk is `phy7` and is slot 0, and the bootable backup disk is `phy6` and is slot 1.

NOTICE: To create a clean backup, Cray recommends shutting down the Cray system before beginning this procedure.

Also, be aware that there may be a considerable load on the SMW while creating the SMW bootable backup drive.

1. Log on to the SMW as `crayadm` and `su` to `root`.

```
crayadm@smw> su -
Password:
smw#
```

2. Standardize the SMW's boot-time drive names with the Linux run-time drive names.

IMPORTANT: If the SMW configuration files on the SMW root drive have been modified already (because this site has completed this step at least once after installing the updated SMW base operating system), skip to step 3 on page 347; otherwise, complete this step to standardize the SMW's boot-time drive names with the Linux run-time drive names.

Set up ordered drives on the R815 SMW.

- a. Identify the installed SMW drive model numbers, serial numbers, and associated Linux device (`/dev`) names.

Execute `smwmapdrives` on the SMW to identify local (internal) drives mounted in the SMW and provide their Linux device (`/dev`) names.

NOTE: Effective with the SMW 7.2.UP00 release, the `smwmapdrives` script was provided both as a separate file in the release and in the base operating system RPM. If running that release or a later one, use the installed version of the script to back up the SMW.

```
smw# smwmapdrives
List of SMW-installed disk drives
-----
Physical slot 0:
/dev/sda
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RDS
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Physical slot 1:
/dev/sdc
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RD7
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Physical slot 2:
/dev/sdd
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RF3
/dev/disk/by-id/scsi-SATA_FUJITSU_MHZ2160_K85DTB227RF3
/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0
Physical slot 3:
/dev/sdb
/dev/disk/by-id/ata-ST9500620NS_9XF0665V
/dev/disk/by-id/scsi-SATA_ST9500620NS_9XF0665V
/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0
Physical slot 4:
NOT INSTALLED
Physical slot 5:
NOT INSTALLED
```

The device names for `by-id` are persistent and will reference the drive, regardless of the slot in which the drive is installed.

`by-path` names reference a physical drive slot only and do not identify the drive installed in that slot. This is the naming used by default for the logging and database drives when the SMW was installed. This `by-path` name is used to specifically install logging and database file systems because the `by-id`

device names refer to the physical drive slots expected to be used for those file systems and are provided as the default examples in the SMW installation configuration process.

The `/dev/sdX` drive names are not persistent; these names can change with each SMW boot and will change if drives are added, removed, or reordered in the SMW slots.

Choose either the `by-id` naming or the `by-path` naming as the site administrative policy for managing the SMW-install disk drives. The following documentation provides the steps necessary to implement this selection on the SMW prior to creating an SMW bootable backup drive.

- b. Back up the following files before proceeding:

```
smw# cp -p /boot/grub/device.map /boot/grub/device.map-YYYYMMDD
smw# cp -p /boot/grub/menu.lst /boot/grub/menu.lst-YYYYMMDD
smw# cp -p /etc/fstab /etc/fstab-YYYYMMDD
```

Cray recommends that `/boot/grub/device.map`, `/etc/fstab` and `/boot/grub/menu.lst` changes use the "by-path" rather than the "by-id" device name because that would allow physically swapping the backup drive into the primary slot when there is a disk failure in the primary disk. If the backup disk is intended as backup only, rather than as a bootable backup, it is acceptable to use either device name.

- c. Edit the grub `device.map` file to reflect physical drive locations.

To provide a direct mapping of the SMW disk drive physical slots to the boot loader (BIOS and `grub`) drive names, the `device.map` mapping file used by `grub` should be replaced. Perform the following steps to install new `device.map` file entries to effect this mapping.

1. Edit the grub `device.map` file.
2. Delete all lines.
3. Enter the following lines into the file. These lines show each drive slot's physical location mapped to its boot-time `hd?` name. Note that `by-id` names should not be used in the `device.map` file.

```
# Dell Rackmount r815 SMW
# grub(8) device mapping for boot-drive identification
# hd? numbers are being mapped to their physical
(hd0) /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-
lun-0
(hd1) /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-
lun-0
(hd2) /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-
lun-0
(hd3) /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-
lun-0
(hd4) /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-
lun-0
(hd5) /dev/disk/by-path/pci-0000:05:00.0-sas-phy2-0x4433221102000000-
lun-0
```

- d. Modify the SMW boot drive `/etc/fstab` file to use `by-id` or `by-path` naming.

Modify the SMW file system mounting configuration file to use SMW disk `by-id` or `by-path` naming. Complete this step to replace any `/dev/sdX` disk partition references.

NOTE: Use the output of the `smwmapdrives` script in step 2.a on page 345 as a reference for drive names.

Edit `/etc/fstab`, replacing drive `/dev/sdX` references with either the `by-id` or `by-path` name's corresponding device name.

When a reference to `/dev/sda1` is being replaced, replace it with the corresponding "partition" file system suffixed with `-part1`. File system partitions for `/dev/sda` are indicated by the numeral appended to the device name; for example, `/dev/sda1` refers to partition 1 on `/dev/sda`.

For example, if the root and swap file systems are currently configured to mount `/dev/sda2`, they should be changed. Using the `by-path` device name from the example in step 2.a on page 345, the `fstab` lines would change from:

```
/dev/sda1 swap          swap          defaults      0 0
/dev/sda2 /                ext3          acl,user_xattr 1 1
```

to:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part1 swap swap defaults      0 0
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-
part2 /      ext3    acl,user_xattr 1 1
```

- e. Modify `/boot/grub/menu.lst` to reflect the `device.map` BIOS/boot-up drive changes for the `sdX` remapping.

The same device name replacement performed on `/etc/fstab` should also be performed on the grub bootloader `/boot/grub/menu.lst` configuration file. All references to `/dev/sdX` devices should be replaced with corresponding `by-path` device names.

- f. Invoke the grub utility to reinstall the SMW boot loader on the primary boot drive.

Once the changes to `device.map`, `fstab`, and `menu.lst` have been completed, the grub bootloader boot blocks must be updated to reflect changes to the device names. Complete this step to update the boot loader on the boot drive.

Invoke the grub utility and reinstall SMW root-drive boot blocks.

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
  lists possible command completions. Anywhere else TAB lists the possible
  completions of a device/filename. ]
grub> root (hd0,1)
      root (hd0,1)
      Filesystem type is ext2fs, partition type 0x83
grub> setup (hd0)
      Checking if "/boot/grub/stage1" exists... yes
      Checking if "/boot/grub/stage2" exists... yes
      Checking if "/boot/grub/e2fs_stage1_5" exists... yes
      Running "embed /boot/grub/e2fs_stage1_5 (hd0)"... 17 sectors \
are embedded. Succeeded
      Running "install /boot/grub/stage1 (hd0) (hd0)1+17 p (hd0,1)/boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
      Done.
grub> quit
```

3. If the backup drive disk partition table already exists and the partition table on the backup drive matches the partition table that is on the primary boot drive, skip this step; otherwise, create the backup drive disk partition table.

In this example, the partition table consists of two slices. Slice 1 is a 4 GB Linux swap partition. Slice 2 is the balance of disk space used for the root file system.

- a. Use the `fdisk` command to display the boot disk partition layout.

```
smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000082

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part1 \
    Boot   Start       End     Blocks  Id System
          63 16771859   8385898+ 82 Linux swap / Solaris
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 \
    Boot   Start       End     Blocks  Id System
          * 16771860 312576704 147902422+ 83 Linux
```

- b. Use the `fdisk` command to configure the bootable backup disk partition layout. Set the bootable backup disk partition layout to match the boot disk partition layout. First, clear all of the old partitions using the `d` command within `fdisk`; next create a Linux swap and a Linux partition; and then write the changes to the disk. For help, type `m` within `fdisk`.

```
smw# fdisk -u /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0

The number of cylinders for this disk is set to 19457.
There is nothing wrong with that, but this is larger than 1024,
and could in certain setups cause problems with:
 1) software that runs at boot time (e.g., old versions of LILO)
 2) booting and partitioning software from other OSs
   (e.g., DOS FDISK, OS/2 FDISK)

Command (m for help): p

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0: 250.0 GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors
Units = sectors of 1 * 512 = 512 bytes
Disk identifier: 0x00000080

    Device
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1 \
    Boot   Start       End     Blocks  Id System
          63 16771859   83828   82 Linux
swap / Solaris
Partition 1 does not end on cylinder boundary.
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 \
    Boot   Start       End     Blocks  Id System
          167719 312581807 156207044+ 83 Linux

Command (m for help): d
Partition number (1-4): 2

Command (m for help): d
Selected partition 1

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 1
```

```

First sector (63-312581807, default 63): (Press the Enter key)
Using default value 63
Last sector, +sectors or +size{K,M,G} (63-312581807, default 312581807): 16771859
Command (m for help): t
Selected partition 1
Hex code (type L to list codes): 82
Changed system type of partition 1 to 82 (Linux swap / Solaris)

Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)
p
Partition number (1-4): 2
First sector (16771860-312581807, default 16771860): (Press the Enter key)
Using default value 16771860
Last sector, +sectors or +size{K,M,G} (16771860-312581807, default 312581807): (Press
the Enter key)
Using default value 312581807

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.

```

- c. Display the boot backup disk partition layout and confirm it matches the `phy7` sector information.

```

smw# fdisk -lu /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0
Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0: 250.0
GB, \
268435456000 bytes
255 heads, 63 sectors/track, 19457 cylinders, total 312581808 sectors

```

4. Initialize the swap device.

```

smw# mkswap /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part1
mkswap: /dev/disk/by-path/pci-0000:05:00.0-sas-phy6:1-0x4433221106000000:0-lun0-part1:
warning: don't erase bootbits sectors
(DOS partition table detected). Use -f to force.
Setting up swapspace version 1, size = 8385892 KiB
no label, UUID=c0ef22ac-b405-4236-855b-e4a09b6e94ed

```

5. Create a new file system on the backup drive root partition by executing the `mkfs` command.

```

smw# mkfs -t ext3 /dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-
part2
mke2fs 1.41.9 (22-Aug-2009)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
9248768 inodes, 36976243 blocks
1848812 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
1129 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
    32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208,
    4096000, 7962624, 11239424, 20480000, 23887872

Writing inode tables:   done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

```

This filesystem will be automatically checked every 37 mounts or 180 days, whichever comes first. Use `tune2fs -c` or `-i` to override.

6. Mount the new backup root file system on `/mnt`.

```
smw# mount \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy6-0x4433221106000000-lun-0-part2 /mnt
```

7. Confirm that the backup root file system is mounted.

```
smw# df
Filesystem            1K-blocks      Used Available Use% Mounted on
/dev/sda2              303528624    6438700 281671544   3% /
udev                  1030332         116   1030216    1% /dev
/dev/sdb2              306128812    195568 290505224   1% /mnt
```

The running root file system device is the one mounted on `/`.

8. Dump the running root file system to the backup drive.

```
smw# cd /mnt
smw# dump 0f - \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy7-0x4433221107000000-lun-0-part2 | restore rf -
DUMP: WARNING: no file `/etc/dumpdates'
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Dumping /dev/sda2 (/) to standard output
DUMP: Label: none
DUMP: Writing 10 Kilobyte records
DUMP: mapping (Pass I) [regular files]
DUMP: mapping (Pass II) [directories]
DUMP: estimated 7898711 blocks.
DUMP: Volume 1 started with block 1 at: Tue Mar 15 13:44:40 2011
DUMP: dumping (Pass III) [directories]
DUMP: dumping (Pass IV) [regular files]
restore: ./lost+found: File exists
DUMP: 79.34% done at 20890 kB/s, finished in 0:01
DUMP: Volume 1 completed at: Tue Mar 15 13:52:13 2011
DUMP: Volume 1 7908080 blocks (7722.73MB)
DUMP: Volume 1 took 0:07:33
DUMP: Volume 1 transfer rate: 17457 kB/s
DUMP: 7908080 blocks (7722.73MB)
DUMP: finished in 453 seconds, throughput 17457 kBytes/sec
DUMP: Date of this level 0 dump: Tue Mar 15 13:43:17 2011
DUMP: Date this dump completed: Tue Mar 15 13:52:13 2011
DUMP: Average transfer rate: 17457 kB/s
DUMP: DUMP IS DONE
```

9. Modify the backup drive's `fstab` and `menu.lst` files to reflect the backup drive's device, replacing the primary drive's device name.

NOTE: This step is necessary only if `by-id` names are used. If `by-path` names are being utilized for the root and swap devices, changes are not necessary; these devices reference physical slots, and the backup drive will be moved to the same physical slot (slot 0) when replacing a failed primary boot drive.

- a. Edit `/mnt/etc/fstab`. Replace the root and swap partitions' `by-id` device names with those used for this backup device, replacing the original disk device name.
For example, change

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2 / ext3 acl,user_xattr
```

to:

```
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part1 swap swap defaults
/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2 / ext3 acl,user_xattr
```

- b. Edit `/mnt/boot/grub/menu.lst`. Replace the `root=` and `resume=` device names with those used for this backup device, replacing the original disk device name.

The `root=` entry normally refers to partition `-part2`, and the `resume=` entry normally refers to partition `-part1`; these partition references must be maintained.

For example, replace the `menu.lst` configuration references of:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RDS-part2
```

with:

```
root=/dev/disk/by-id/ata-FUJITSU_MHZ2160BK_G2_K85DTB227RD7-part2
```

or similarly with the `by-id` device names, if those are preferred.

Replace the `resume=` references similarly.

10. Install the `grub` boot loader. To make the backup drive bootable, reinstall the `grub` boot facility on that drive.



CAUTION: Although all of the disks connected to the SMW are available to the system, `grub` detects only the first 16 devices. Therefore, if a disk is added to the SMW after the SMW is connected to the boot RAID, it is advisable to reboot the SMW before continuing.

- a. Create a unique file on the backup drive to be used to identify that drive to `grub` boot facility.

```
smw# cd /
smw# touch /mnt/THIS_IS_6
```

- b. Invoke the `grub` boot utility. Within the `grub` boot utility:

1. Execute the `find` command to locate the drive designation that `grub` uses.
2. Select the drive to which the boot blocks will be installed with the `root` command.
3. Use the `setup` command to set up and install the `grub` boot blocks on that drive. The Linux `grub` utility and boot system *always* refer to drives as `hd`, regardless of the actual type of drives. For example:

```
smw# grub --no-curses
GNU GRUB version 0.97 (640K lower / 3072K upper memory)
[ Minimal BASH-like line editing is supported. For the first word, TAB
lists possible command completions. Anywhere else TAB lists the
possible
completions of a device/filename. ]
grub> find /THIS_IS_6
(hd2,1)
grub> root (hd2,1)
root (hd2,1)
Filesystem type is ext2fs, partition type 0x83
grub> setup (hd2)
Checking if "/boot/grub/stage1" exists... yes
Checking if "/boot/grub/stage2" exists... yes
Checking if "/boot/grub/e2fs_stage1_5" exists... yes
Running "embed /boot/grub/e2fs_stage1_5 (hd2)"... 17 sectors are
```

```

embedded.
succeeded
Running "install /boot/grub/stage1 (hd2) (hd2)1+17 p (hd2,1) /boot/grub/
stage2 \
/boot/grub/menu.lst"... succeeded
Done.
grub> quit

```

IMPORTANT: For R815 SMWs, grub recreates `device.map` with the short names, not the persistent names. Do not trust the `/dev/sdx` names. Always use `find` when executing grub because it is possible that `grub root` may not be `hd2` the next time grub is executed.

11. Unmount the backup root partition.

```
smw# umount /mnt
```

The drive is now bootable once plugged in or cabled as the primary drive.

Update the SMW Software and Configuration

1. Open a terminal window, and `su` to `root`.

```
crayadm@smw> su - root
smw#
```

2. Mount the release media by using one of the following commands, depending on the media type.

- To install the update package from disk, place the SMW 7.2UP04 Software DVD in the CD/DVD drive and mount it.

```
smw# mount /dev/cdrom /media/cdrom
```

- To install the update package from disk images instead of from the DVD, copy the files to a directory such as `/tmp/SMW_version` on the SMW and then substitute this path for `/media/cdrom` in subsequent instructions.
- To install the update package using the ISO image, in the current directory execute the following command with the file name of the `smw-image` ISO image for the update being installed. For example:

```
smw# mount -o loop,ro smw-image-7.2.0-1.0702.37336.662-1.iso /media/cdrom
```

3. If `postfix` is not configured on the SMW, skip this step:

To prevent the `master.cf` and `main.cf` `postfix` configuration files from being recreated during software updates or fixes, ensure the following setting in the `/etc/sysconfig/mail` file on the SMW is set to "no":

```
MAIL_CREATE_CONFIG="no"
```

4. To see what SMW software will be updated in this new release, execute these commands prior to doing the update. This information is gathered, displayed, and contained in the log files during the `SMWinstall` process.
 - a. Check security and recommended updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GS
```

- b. Check Cray software updates.

```
smw# /media/cdrom/CRAYSMWinstall.sh -GV
```

5. Create a new copy of the `SMWinstall.conf` configuration file and modify the new copy of the `SMWinstall.conf` file with site-specific requirements. Only `root` can modify the `SMWinstall.conf` configuration file. The `SMWinstall.conf` configuration file is created during the installation process by copying the `SMWinstall.conf` template from the distribution media. By default, the SMW configuration file is placed in `/home/crayadm/SMWinstall.conf`.

```
smw# cp /media/cdrom/SMWinstall.conf /home/crayadm
smw# chmod 644 /home/crayadm/SMWinstall.conf
smw# vi /home/crayadm/SMWinstall.conf
```

IMPORTANT: For an SMW HA system, define this variable as if the system was a standalone SMW. The shared storage for the SMW HSS database will be configured later in the SMW HA configuration process.

For a description of the contents of the `SMWinstall.conf` file, see the `SMWinstall.conf(5)` man page.

6. Update the software with `SMWinstall`. `SMWinstall` checks for any inconsistency between the system and the `SMWinstall.conf` file settings, prompts for the root MySQL database password, and stores its log files in `/var/adm/cray/logs`.

```
smw# /media/cdrom/SMWinstall
...
```

```
Please enter your root DB password:
Please confirm your root DB password:
Password confirmed.
```

When `SMWinstall` finishes, it will suggest a reboot of the SMW.

7. If necessary, restore the locally modified versions of the following files.
 - a. If the installation site had locally modified versions of `/etc/syslog-ng/syslog-ng.conf` or `/etc/rsyslog.conf` before this SMW update, restore the local modifications to these files. During the upgrade procedure, the old files are saved in `/etc/syslog-ng/syslog-ng.conf-YYYYMMDDhhmm` and `/etc/rsyslog.conf.rpmsave`.
 - b. If the installation site had local modifications to `/opt/cray/hss/default/etc/auto.xtshutdown` before this SMW update, restore the local modifications to this file.
 - c. If the installation site had power management profiles, copy the files that were backed up from the `/opt/cray/hss/7.1.0/pm/profiles` directory into the new `/opt/cray/hss/default/pm/profiles` directory. See *Monitoring and Managing Power Consumption on the Cray XC30 System (S-0043)* for instructions on validating the restored profiles.
 - d. If the installation site had local modifications to `/opt/cray/hss/default/etc/sedc_srv.ini` before this SMW update, locate the destination of this symbolic link (see [Prepare to Upgrade or Update SMW Software](#) on page 245), compare the content of the local file to the distributed version of the file, and update the local file appropriately.
8. If the installation site has an `/opt/cray/hss/default/etc/xtdiscover.ini` file, the SMW update process does not overwrite an existing `xtdiscover.ini` file; the new version is named `xtdiscover.ini.dist`. Compare the content of the new `xtdiscover.ini.dist` with the original `xtdiscover.ini` file, and update the `xtdiscover.ini` file appropriately.

NOTE: If the `xtdiscover.ini` file does not exist, then the `xtdiscover.ini.dist` file is copied to the `xtdiscover.ini` file.

9. Unmount the SMW 7.2UP04 Software media.

```
smw# umount /media/cdrom
```

10. If using the update disk, eject the SMW 7.2UP04 Software DVD.

```
smw# eject
```

11. Reboot the SMW.

```
smw# reboot
```

For a Cray XE or Cray XK system, continue to [Update the L0 and L1 Firmware](#).

Update the Controller Boot Image for the Passive SMW

Prerequisites

Before using this procedure, ensure that the SMW software and configuration has been updated on the passive SMW.

On the passive SMW, update only the controller boot image. Do not discover the hardware or update the firmware; these steps were done on the active SMW.

1. If necessary, log in as `root` to the passive SMW.
2. Update the controller boot image.
The version used in the command argument for `hss_make_default_initrd` should match that of the version specified in the `lsb-cray-hss` line in the output from the `crms-release` file. This directory will not exist until the `hss_make_default_initrd` command completes.

```
smw# cat /opt/cray/hss/default/etc/crms-release
HSS-CRMS Mon Sep 14 00:57:20 CDT 2015 on hssbld0 by bwdev
lsb-cray-hss-7.2.0-1.0702.37336.662
smw# hss_make_default_initrd /opt/cray/hss-images/master/7.2.0-1.0702.37336.662
::: Verifying base RPM list to the manifest
::: Installing filesystem hierarchy
::: Installing busybox
::: Installing base RPMs
::: Installing ssh
::: Removing /etc/securetty installed by the pam RPM
::: Installing sshfs
::: Installing rsh
::: Modifying /etc/pam.d/rlogin to remove securetty checking
::: Modifying /etc/pam.d/rsh to remove rhosts and nologin checking (Bug #779466)
::: Installing rsync
::: Installing atftp
::: Installing tcpdump
::: Installing ethtool
::: Installing syslog-ng
::: Installing logrotate
::: Installing ntp/ntpd
```

```

::: Installing strace
::: Installing screen
::: Installing minicom
::: Installing ppp
::: Installing mtd-utils
::: Installing /init
::: Installing file.rpm
::: Installing libgmodule
::: Installing Midnight Commander
::: Installing cray-viper
::: Installing spread
::: Installing coreboot-utils
::: Clearing init.d to be replaced by cray-hss32-filesystem
::: Creating initial etc files needed for root creation
::: Installing Cray kernel
::: Installing latest Cray kernel modules
::: Clearing select /boot items
::: Installing boot-parameters
::: Installing cray-hss32-scripts
::: Installing lsb-cray-hss-controllers
::: Installing cray-libconfig
::: Installing cray-bdm
::: Installing cray-play_xsvf
::: Removing unwanted files from the root

=====
The new initrd hierarchy is now in /opt/cray/hss-images/master/
7.2.0-1.0702.37336.662.

Running hssclone.
Image Clone Complete: /opt/cray/hss-images/image-7.2.0-1.0702.37336.662
Running hsspackage.
copying image
copying modules
running depmod
creating load file: /opt/cray/hss-images/default/HSS32.load
compressing initrd.img
Creating pxelinux.0 symlink
Running hssbootlink.
linking /opt/cray/hss-images/default/HSS32/bzImage-3.0.76-0.11.1_1.0702.8867-
cray_hss32 /opt/tftpboot/bzImage
linking /opt/cray/hss-images/default/HSS32/parameters /opt/tftpboot/
pxelinux.cfg/default
linking /opt/cray/hss-images/default/HSS32/initrd.img /opt/tftpboot/initrd.img

```

Finish Updating the SMW Software on the Passive SMW

After updating the SMW software on the passive SMW, use this procedure to finish the update.

1. Reboot the passive SMW.
2. Log into the passive SMW as `root`.

Finish the SMW Update

Prerequisites

Before beginning this procedure, log into both SMWs as `root`.

Use this procedure to finish the SMW update. The examples in this procedure assume that `smw1` was the active SMW at the start of the update and is currently the passive SMW.

1. On the second SMW (`smw2`), turn off maintenance mode.

```
smw2:~ # crm configure property maintenance-mode=false 2> /dev/null
```

2. On `smw1`, start file synchronizing.

```
smw1:~ # crm resource start fsync
```

3. Check that the `rsms`, `dbMonitor`, and `mysql` services are disabled on both SMWs. These services must be off when the SMWs reboot.

```
smw1:~ # chkconfig -list rsms dbMonitor mysql
rsms          0:off 1:off 2:off 3:off 4:off 5:off 6:off
dbMonitor     0:off 1:off 2:off 3:off 4:off 5:off 6:off
mysql         0:off 1:off 2:off 3:off 4:off 5:off 6:off

smw2:~ # chkconfig -list rsms dbMonitor mysql
rsms          0:off 1:off 2:off 3:off 4:off 5:off 6:off
dbMonitor     0:off 1:off 2:off 3:off 4:off 5:off 6:off
mysql         0:off 1:off 2:off 3:off 4:off 5:off 6:off
```

If any of these services are on, use the following commands to turn them off.

```
smw1:~ # chkconfig rsms off
smw1:~ # chkconfig dbMonitor off
smw1:~ # chkconfig mysql off

smw2:~ # chkconfig rsms off
smw2:~ # chkconfig dbMonitor off
smw2:~ # chkconfig mysql off
```

4. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

5. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

6. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                    Started smw1
stonith-1      (stonith:external/ipmi):       Started smw2
stonith-2      (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir     (ocf::heartbeat:Filesystem):    Started smw1
  md-fs       (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
```

```
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

- From either SMW, execute the `clear_failcounts` command to clean up any SMW HA resource errors.

```
smw1:~ # clear_failcounts
```

- If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, restart the VNC server.
- Return the `postgresql` service to its pre-update state (either on or off), as recorded in [Prepare the SMW HA System for an SMW Upgrade](#) on page 244.

```
smw1:~ # chkconfig postgresql state
```

Update SMW HA Software

Prerequisites

Before you start this procedure:

- Ensure that both SMWs are running the required operating system and SMW software.
- Check that FN-6029 has been installed.

For more information, see [Before You Start an SMW HA Update](#) on page 325.

To update the SMW HA software, update the active SMW first, then update the other SMW.

- Log in as `root` to the first SMW.

```
workstation> ssh root@smw1
```

- In a separate terminal session, log in as `root` to the second SMW.

```
workstation> ssh root@smw2
```

- Find the active SMW by determining where the SMW HA cluster resources are running (such as the `hss-daemons` resource).

```
smw1:~ # crm_mon -1 | grep hss-daemons
hss-daemons (lsb:rsms): Started smw1
```

NOTE: The examples in this procedure assume that `smw1` is currently the active SMW.

- Stop the OpenAIS service on both SMWs simultaneously.

```
smw1:~ # rcopenais stop
```

```
smw2:~ # rcopenais stop
```

5. Update the SMW HA software on the active SMW (`smw1`).

a. Mount the Cray SMW HA release media on the SMW.

- If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw1:~ # mount /dev/cdrom /media/cdrom
```

- If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

IMPORTANT: The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

The ISO file name depends on the release number and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For *path*, substitute the actual path to the ISO on the system.

```
smw1:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

b. Navigate to the `/media/cdrom` directory.

```
smw1:~ # cd /media/cdrom
```

c. Install the Cray SMW HA release software on the SMW.

```
smw1:~ # ./SMWHAinstall -v
```

d. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.

e. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw1:~ # cd
smw1:~ # umount /media/cdrom
smw1:~ # eject
```

6. Update the SMW HA software on the other SMW (`smw2`).

a. Mount the Cray SMW HA release media on the SMW.

- If you have the release media on DVD, place the Cray SMW HA DVD into the DVD drive and mount it to `/media/cdrom`.

```
smw2:~ # mount /dev/cdrom /media/cdrom
```

- If you have an electronic version of the release media, mount the Cray SMW HA ISO to `/media/cdrom`.

IMPORTANT: The ISO must reside on a local disk (such as `/tmp`), not on a shared file system on the boot RAID. For example, do not use a subdirectory in `/home`.

The ISO file name depends on the release number and installer version. The following command shows the generic ISO name `smw-SLEHA11SP3xnn.iso`, where `xnn` is the installer version (for example, `smw-SLEHA11SP3b06.iso`). Substitute the actual ISO file name as specified in the SMW HA release information.

For `path`, substitute the actual path to the ISO on the system.

```
smw2:~ # mount -o loop,ro /path/smw-SLEHA11SP3xnn.iso /media/cdrom
```

- b. Navigate to the `/media/cdrom` directory.

```
smw2:~ # cd /media/cdrom
```

- c. Install the Cray SMW HA release software on the SMW.

```
smw2:~ # ./SMWHAinstall -v
```

- d. Examine the initial output from `SMWHAinstall` and check the log file, if necessary. `SMWHAinstall` creates a log file in `/var/adm/cray/logs/SMWHAinstall.timestamp.log`.

- e. Navigate out of the `/media/cdrom` directory and unmount the SMW HA release media. If you are using a physical DVD, also eject the DVD.

```
smw2:~ # cd
smw2:~ # umount /media/cdrom
smw2:~ # eject
```

7. Start the OpenAIS service on both SMWs simultaneously.

```
smw1:~ # rcopenais start
```

```
smw2:~ # rcopenais start
```

8. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

9. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

10. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
```

```
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons   (lsb:rsms):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):   Started smw1
  cray-syslog (lsb:cray-syslog):          Started smw1
  homedir     (ocf::heartbeat:Filesystem):   Started smw1
  md-fs       (ocf::heartbeat:Filesystem):   Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):   Started smw1
  postgresql  (lsb:postgresql):            Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

11. Update the SMW HA cluster configuration.

- a. If you are using Virtual Network Computing (VNC) software to enable remote access to the SMW, shut down the VNC server.

```
smw1:~ # /etc/init.d/vnc stop
```

- b. On the active SMW, change to the directory containing the SMWHAconfig command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

- c. Load the `ha-smw` module.

```
smw1:~ # module load ha-smw
```

- d. On the active SMW, execute the `SMWHAconfig` command with the `--update` option.

```
smw1:~ # ./SMWHAconfig --update
```

- e. When `SMWHAconfig` runs, it may prompt for the virtual host name if the system is being updated from an older version of the release. If so, enter the virtual host name for the SMW HA cluster.

- f. If necessary, examine the log file. `SMWHAconfig` creates a log file in `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out`.

- g. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

- h. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

- i. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
```

```

ClusterIP4      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):  Started smw1
Notification    (ocf::heartbeat:MailTo):      Started smw1
dhcpd           (lsb:dhcpd):      Started smw1
fsync           (ocf::smw:fsync):      Started smw1
hss-daemons     (lsb:rsmc):      Started smw1
stonith-1       (stonith:external/ipmi):      Started smw2
stonith-2       (stonith:external/ipmi):      Started smw1
Resource Group: HSSGroup
  ml-fs         (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog   (lsb:cray-syslog):      Started smw1
  homedir       (ocf::heartbeat:Filesystem):    Started smw1
  md-fs         (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs         (ocf::heartbeat:Filesystem):    Started smw1
  postgresql    (lsb:postgresql):      Started smw1
  mysqld        (ocf::heartbeat:mysql):    Started smw1

```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- j. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- k. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```

smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

- l. If necessary, restart the VNC server.
12. If you have made local changes to the list of synchronized files and directories in `/etc/csync2/csync2_cray.cfg`, restore the local changes to the updated file.

The installation procedure saves local changes in the file `/etc/csync2/csync2_cray.cfg.sav`. You must copy these changes into `csync2_cray.cfg`.

- a. On `smw1`, navigate to the `/etc/csync2` directory.

- b. Edit the files `csync2_cray.cfg` and `csync2_cray.cfg.sav`.

NOTE: You can ignore the generic host entries near the top of the file. The `SMWHAconfig` command will restore site-specific host entries later in this procedure.

- c. Locate the `group user_group` section in `csync2_cray.cfg.sav`, and copy the include and exclude lines into `csync2_cray.cfg`.
- d. Save your changes to `csync2_cray.cfg` and exit the editor for both files.

13. From either SMW, execute the `clear_failcounts` command to clean up any SMW HA resource errors.

```
smw1:~ # clear_failcounts
```

14. Display the cluster status and verify that each resource has been started.

```
smw1:~ # crm_mon -1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):      Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1
fsync (ocf::smw:fsync):      Started smw1
hss-daemons (lsb:rsms):      Started smw1
stonith-1     (stonith:external/ipmi):       Started smw2
stonith-2     (stonith:external/ipmi):       Started smw1
Resource Group: HSSGroup
  ml-fs       (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog (lsb:cray-syslog):           Started smw1
  homedir    (ocf::heartbeat:Filesystem):    Started smw1
  md-fs      (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs      (ocf::heartbeat:Filesystem):    Started smw1
  postgresql (lsb:postgresql):         Started smw1
  mysqld     (ocf::heartbeat:mysql):         Started smw1
```

`crm_mon` may display different resource names, group names, or resource order on the system.

15. If any resource status is `Stopped`, execute the `clean_resources` command.

```
smw1:~ # clean_resources
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command.

16. Display the resource failcount data. All failcounts should be zero.

```
smw1:~# show_failcounts
node=smw1 scope=status name=fail-count-stonith-1 value=0
node=smw1 scope=status name=fail-count-stonith-2 value=0
node=smw1 scope=status name=fail-count-dhcpd value=0
node=smw1 scope=status name=fail-count-cray-syslog value=0
...
```

If there is a problem with the update, see [Restore a Previous SMW HA Configuration](#) on page 399.

Update CLE Software on the SMW HA System

When updating the SMW HA system, Cray recommends updating the CLE software after you have updated the SMW and SMW HA software. If necessary, you can update the CLE software before you update the SMW and SMW HA software.

Use the following procedures to upgrade the CLE software on the SMW HA system.

1. Update CLE software on the active SMW.
2. Copy boot images and CLE install directory to the passive SMW.
3. Update CLE software on the passive SMW.
4. Finish the CLE update.

Upgrade CLE Software on the Active SMW

Use the following procedures to upgrade the CLE software on the active SMW.

Before Starting the Update or Upgrade Process

Perform the following tasks before you install the CLE release package.

- Read the *README* file provided with the release for any installation-related requirements and corrections to this installation guide.
- Additional installation information may also be included in the following documents:
CLE 5.2.UP04 Release Errata, Limitations for CLE 5.2.UP04, Cray Linux Environment (CLE) Software Release Overview (S-2425), and Cray Linux Environment (CLE) Software Release Overview Supplement (S-2497).
- Verify that your System Management Workstation (SMW) is running Cray SMW Release 7.2.UP04 or later. You must install the SMW 7.2.UP04 release or later on your SMW before installing the CLE 5.2.UP04 release. If a specific SMW update package is required for your installation, that information is documented in the *README* file provided with the CLE 5.2.UP04 release. Type the following command to determine the HSS/SMW version:

```
crayadm@smw:~> cat /opt/cray/hss/default/etc/smw-release
7.2.UP04
```

Back Up the Current Software

Before you install the release package, back up the contents of the system set being updated or upgraded. Use the `xthotbackup` command to back up one system set to a second system set. For more information about using system sets, see [About System Set Configuration in `/etc/sysset.conf`](#) on page 88 and the `sysset.conf(5)` man page.

By default, `xthotbackup` copies only the boot node root and shared root file systems. Specify the `-a` option to copy all file systems in the system set (except for swap and Lustre) or specify the `-f` option to select a customized set of file system functions. The `-b` option makes the backup or destination system set bootable by changing the appropriate boot node and service node entries in `/etc/fstab`. Doing a live backup (`xthotbackup -L`) can reduce the amount of time a CLE system is unavailable to the user community for the CLE backup and software upgrade process. For more information, see the `xthotbackup(8)` man page.

Back Up Current Software

Use the `xthotbackup` command to copy the disk partitions in one system set to a backup system set.



WARNING: If the source system set is booted, you should use the `xthotbackup -L` option. If not using the `xthotbackup -L` option, neither the source system set nor the destination system set should be used by a booted CLE system. Running `xthotbackup` with a booted system set or partition could cause data corruption.

1. If the Cray system is booted, use your site-specific procedures to shut down the system. For example, to shutdown using an automation file:

```
crayadm@smw:~> xtbootsys -s last -a auto.xtshutdown
```

For more information about using automation files see the `xtbootsys(8)` man page.

Although not the preferred method, alternatively execute these commands as `root` from the boot node to shutdown your system.

```
boot:~ # xtshutdown -y
boot:~ # shutdown -h now;exit
```

2. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

3. Run the `xthotbackup` command to copy from the source system set to the backup or destination system set. For example, if `BLUE` is the label for the source system set and `GREEN` is the label for the backup system set, execute the following command as `root`:

```
smw:~ # xthotbackup -a -b BLUE GREEN
```

NOTE: The `-a` option specifies all file system functions in the system set (except swap and Lustre). To specify a site-specific set of functions, use the `-f` option.

`xthotbackup` does not copy the swap partition for the boot node, however, if the `-b` option is specified, `mkswap` is invoked on the swap partition for the boot node in the destination system set to prepare a swap partition.

For more information, see the `xthotbackup(8)` man page.

Back Up Current Software Using xthotbackup -L

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Run the `xthotbackup` command to copy a booted system from the source system set to the backup or destination system set. For example, if `BLUE` is the label for the source system set and `GREEN` is the label for the backup system set, execute the following command:

```
smw:~ # xthotbackup -L -a -b BLUE GREEN
```

NOTE: The `-L` option will connect to the boot node (or other nodes that mount the file systems in the source system set) to backup the file systems in the destination system set. The `-a` option specifies all file system functions in the system set (except swap and Lustre). To specify a site-specific set of functions, use the `-f` option.

The `xthotbackup` command does not copy the swap partition for the boot node. However, if the `-b` option is specified, `mkswap` is invoked on the swap partition for the boot node in the destination system set to prepare a swap partition.

You are now ready to begin installing the software release package.

Update CLE Software

To use the procedures in this chapter, a Cray system must be running CLE 5.2.UP00 to update the CLE software to a newer version of CLE 5.2.

Before You Begin

All upgrades, updates, and configuration changes are installed from the SMW to the `bootroot`, `sharedroot`, and (if applicable) the persistent `/var` file systems before booting the upgraded file systems. These file systems are mounted and modified during the procedure to install the release package.

An update or upgrade release package can be installed to an alternative root location if a system is configured to have more than one system set. A significant portion of the upgrade work can be done without using dedicated time if your Cray system is booted from a different system set. For example, if your system is running on the `BLUE` system set, and the `GREEN` system set is a backup of `BLUE`, then you can perform a large amount of the CLE upgrade process on the `GREEN` system set while the system is booted, thus reducing the amount of system downtime during upgrades/updates. These instructions will inform you when dedicated time is required. The `/etc/sysset.conf` file describes which devices and disk partitions on the boot RAID are used for which system sets. For more information, see [About System Set Configuration in /etc/sysset.conf](#) on page 88 and the `sysset.conf(5)` man page.

If you are updating or upgrading a system set that is not running, you do not need to shut down your Cray system before you install the release package.



WARNING: If you are updating or upgrading a system set that is running, you must shut down your Cray system before installing the release package. For more information about system sets and system startup and shutdown procedures, see *Managing System Software for the Cray Linux Environment (S-2393)*.

If the persistent `/var` file system is shared between multiple system sets, you must verify that it is not mounted on the Cray system before you install the release package.

Install CLE Release Software on the SMW

Two DVDs are provided to install the CLE 5.2 release on a Cray system. The first is labeled Cray CLE 5.2.UPnn Software and contains software specific to Cray systems. Optionally, you may have an ISO image called `xc-sles11sp3-5.2.55d05.iso`, where `5.2.55` indicates the CLE release build level, and `d05` indicates the installer version.

The second DVD is labeled `CentOS-6.5-x86_64-bin-DVD1.iso` and contains the CentOS 6.5 base operating system for CLE direct-attached Lustre (DAL) nodes.

Copy the Software to the SMW

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Mount the release media by using one of the following commands, depending on your media type.

If installing the release package from disk, place the Cray CLE 5.2.UPnn Software DVD in the CD/DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the release media using the ISO image, execute the following command, where `xc-sles11sp3-5.2.55d05.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro xc-sles11sp3-5.2.55d05.iso /media/cdrom
```

3. Copy all files to a directory on the SMW in `/home/crayadm/install.xtrel`, where `xtrel` is a site-determined name specific to the release being installed. For example:

```
smw:~# mkdir /home/crayadm/install.5.2.55
smw:~# cp -pr /media/cdrom/* /home/crayadm/install.5.2.55
```

4. Unmount the Cray CLE 5.2.UPnn Software media.

```
smw:~# umount /media/cdrom
```

5. For upgrading from CLE 5.1 or CLE 4.2 to CLE 5.2, you must mount the SLES 11 SP3 base media. Insert the Cray-CLEbase11sp3 DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the base operating system media using the ISO image, execute the following command, where `Cray-CLEbase11SP3-yyyymmdd.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro Cray-CLEbase11SP3-yyyymmdd.iso /media/cdrom
```

6. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` image.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Install CLE on the SMW

1. As `root`, execute the `CRAYCLEinstall.sh` installation script to upgrade the Cray CLE software on the SMW.

```
smw:~# /home/crayadm/install.5.2.56/CRAYCLEinstall.sh \
-m /home/crayadm/install.5.2.56 -u -v -w
```

2. At the prompt `Do you wish to continue?`, type `y` and press `Enter`.

The output of the installation script is displayed to the console. If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

Preparing the Configuration for Software Installation

You may need to update the `CLEinstall.conf` configuration file. The `CLEinstall.conf` file that was created during the first installation of this system can be used during an installation to the alternative root location. For a description of the contents of this file, see [About Installation Configuration Files](#) on page 74 or the `CLEinstall.conf(5)` man page.

Based on the settings you choose in the `CLEinstall.conf` file, the `CLEinstall` program then updates other configuration files. A template `CLEinstall.conf` is provided on the distribution media. Your site-specific copy is located in the installation directory from the previous installation; for example `/home/crayadm/install.5.2.74/CLEinstall.conf`.



WARNING: Any configuration data which is in `CLEinstall.conf` that was manually changed on a system after the last software update must be kept up to date before running `CLEinstall` for an upgrade or an update. Doing so will prevent spending much time tracking down problems that could have been avoided.

During update and upgrade installations, the `/opt/cray/hss/default/etc/auto.xtshutdown` automated shut-down file is overwritten by the newer shut-down file that corresponds to the update/upgrade release. The old shut-down file will be saved as `/opt/cray/hss/default/etc/auto.xtshutdown.rpmsave`. If your site has made local changes to the autofile, you will need to review the changes and reapply them to the new file following the update.

NOTE: If problems with the hosts file are detected after the update or upgrade, you may need to use the copies of `/etc/hosts` that `CLEinstall` saves on `bootroot` and `/opt/xt-images/templates/default/etc` with `hosts.preinstall.$$` and `hosts.postinstall.$$`.

Prepare the `CLEinstall.conf` Configuration File

1. If you have an existing `CLEinstall.conf` file, use the `diff` command to compare it to the template in `/home/crayadm/install.xtre1`. For example:

```
smw:~# diff /home/crayadm/install.5.2.56/CLEinstall.conf \
/home/crayadm/install.5.2.74/CLEinstall.conf
21c21
< xhostname=mycray
---
> xhostname=crayhostname
24c24
< node_class_login_hostname=mycray
---
```

```
> node_class_login_hostname=crayhostname
smw:~ #
```

NOTE: The CLEinstall program generates INFO messages suggesting that you remove deprecated parameters from your local CLEinstall.conf file.

2. Edit the CLEinstall.conf file in the temporary directory `/home/crayadm/install.xtre1` and make necessary changes to enable any new features you are configuring for the first time with this system software upgrade.

NOTE: The CLEinstall program checks that the `/etc/opt/cray/sdb/node_classes` file and the `node_class[*]` parameters in CLEinstall.conf agree. If you made changes to `/etc/opt/cray/sdb/node_classes` since your last CLE software installation or upgrade, make the same changes to CLEinstall.conf.

```
smw:~# cp -p /home/crayadm/install.5.2.56/CLEinstall.conf \
/home/crayadm/install.5.2.56/CLEinstall.conf.save
smw:~# chmod 644 /home/crayadm/install.5.2.56/CLEinstall.conf
smw:~# vi /home/crayadm/install.5.2.56/CLEinstall.conf
```

For a complete description of the contents of this file, see [About Installation Configuration Files](#) on page 74.

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

Run the CLEinstall Installation Program

The CLEinstall installation program upgrades the CLE software for your configuration by using information in the CLEinstall.conf and sysset.conf configuration files.

IMPORTANT: CLEinstall modifies Cray system entries in `/etc/hosts` each time you update or upgrade your CLE software. For additional information, see [Maintain Node Class Settings and Hostname Aliases](#) on page 76.

If the update or upgrade you are applying modifies configuration information in the `alps.conf` file, your existing `alps.conf` parameters will be automatically merged into the new file and your original file will be saved (in the same directory) as `alps.conf.unmerged`. If you experience problems with ALPS immediately following an update or upgrade, you can replace `alps.conf` with `alps.conf.unmerged` and execute `/etc/init.d/alps restart` on the boot and SDB nodes to restore your original configuration.

During a CLE update or upgrade, CLEinstall disables the execution bits of all scripts in the `/etc/cron.hourly`, `/etc/cron.daily`, `/etc/cron.weekly`, and `/etc/cron.monthly` directories of the bootroot and default view of the shared root with a `chmod ugo-x` command. If there are site-specific cron scripts in these directories, you will need to re-enable the execute permission on them after performing a CLE update or upgrade. Any scripts in these directories which have been node-specialized or class-specialized via `xtopview` will not be changed by the CLE update or upgrade. Only the bootroot and the default view of the shared root will be modified.

The following CLEinstall options are required or recommended for this type of installation:

`--upgrade`

Specify that this is an update or upgrade rather than a full system installation.

`--label=system_set_label`

Specify the system set that you are using to install the release.

`--XTrrelease=release_number`

Specify the target CLE release and build level that you are upgrading to, for example 5.2.55.

--CLEmedia=directory

Specify the directory on the SMW where you copied the CLE software media. For example, `/home/crayadm/install.release_number`.

--configfile=CLEinstall_configuration_file

Specify the path to the `CLEinstall.conf` file that you edited in [Preparing the CLEinstall.conf configuration file](#).

--Centosmedia=directory

Specify the directory where the CentOS software media has been mounted. The `--Centosmedia` option is required when installing or upgrading CLE with direct-attached Lustre (DAL). For example, the CentOS image mount point could be `/media/Centosbase`.

For a full description of the `CLEinstall` command options and arguments, see [Run the CLEinstall Program](#) on page 94 or the `CLEinstall(8)` man page.

Run CLEinstall

1. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` image if it is not already mounted.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Also include the `--Centosmedia=directory` option when invoking `CLEinstall`. In this example, the option is `--Centosmedia=/media/Centosbase`.

2. Invoke the `CLEinstall` program on the SMW. `CLEinstall` is located in the directory you created in [Copy the Software to the SMW](#).

```
smw:~# /home/crayadm/install.5.2.55/CLEinstall --upgrade \
--label=system_set_label --XTrelease=5.2.55 \
--configfile=/home/crayadm/install.5.2.55/CLEinstall.conf \
--CLEmedia=/home/crayadm/install.5.2.55 \
--Basemedia=/media/cdrom
```

3. Examine the initial messages directed to standard output. Log files are created in `/var/adm/cray/logs` and named by using a timestamp that indicates when the install script began executing. For example:

```
08:57:48 Installation output will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.stdout.log
08:57:48 Installation errors (stderr) will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.stderr.log
08:57:48 Installation debugging messages will be captured in /var/adm/cray/logs/\
CLEinstall.p3.20140911085748.CLE52-P3.debug.log
```

The naming conventions of these logs are:

`CLEinstall.p#.YYYYMMDDhhmmss.$LABEL.logtype.log`

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format.

`$LABEL` is the system set label (in the example above, `CLE52-P3`).

`logtype` is `stdout` (standard output), `stderr` (standard error), or `debug`.

Also, log files are created in `/var/adm/cray/logs` each time `CLEinstall` calls `CRAYCLEinstall.sh`. For example:

```
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.01-B.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.02-B.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.03-B.log
.
.
.
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.17-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.18-S.log
CRAYCLEinstall.sh.p3.20140911085748.CLE52-P3.19-S.log
```

The naming conventions of these logs are:

```
CRAYCLEinstall.sh.p#.YYYYMMDDhhmmss.$LABEL.sequence#-root.log
```

where

`p#` is the partition number specified in the `CLEinstall.conf` file.

`YYYYMMDDhhmmss` is the timestamp in year, month, day, hour, minute, and second format. This is the same timestamp used for the log files of the `CLEinstall` program instance that called `CRAYCLEinstall.sh`.

`$LABEL` is the system set label.

`sequence#` is an increasing count that specifies each invocation of `CRAYCLEinstall.sh` by `CLEinstall`.

`root` is either `B` (bootroot) or `S` (sharedroot), specifying the root modified by the `CRAYCLEinstall.sh` call.

4. `CLEinstall` validates `sysset.conf` and `CLEinstall.conf` configuration settings and then confirms the expected status of your boot node and file systems.

Confirm that the installation is proceeding as expected, respond to warnings and prompts, and resolve any issues. For example:

- If you are installing to a system set that is not running, and you did not shut down your Cray system, respond to the following warning and prompt:

```
WARNING: Your bootnode is booted. Please confirm that the
system set you are intending to update is not booted.
Do you wish to proceed?[n]:
```



WARNING: If the boot node has a file system mounted and `CLEinstall` on the SMW creates a new file system on that disk partition, the running system will be corrupted.

- If you have configured file systems that are shared between two system sets, respond to the following prompt to confirm creation of new file systems:

```
09:21:24 INFO: The PERSISTENT_VAR disk function for the LABEL system set is marked shared.
09:21:24 INFO: The /dev/sdrl disk partition will be mounted on the SMW for PERSISTENT_VAR
disk function. Confirm that it is not mounted on any nodes in a running XT
system before continuing.
Do you wish to proceed?[n]:y
```

- If the `node_classidx` parameters do not match the existing `/etc/opt/cray/sdb/node_classes` file, you are asked to confirm that your hardware configuration has changed. If your hardware has not changed, abort `CLEinstall` and correct the node class configuration in `CLEinstall.conf` and/or the `node_classes` file. Respond to the following warning and prompt:

```
09:21:41 INFO: There are 5 WARNINGS about discrepancies between CLEinstall.conf
and /etc/opt/cray/sdb/node_classes
09:21:41 INFO: If you ARE adding service nodes, then you may proceed and CLEinstall
```

```
will adjust the /etc/opt/cray/sdb/node_classes file to match the settings in CLEinstall.conf and may remove some node-specialized files from the shared root specialized /etc.
09:21:41 INFO: If you ARE NOT adding service nodes, then stop CLEinstall now to correct the problem.
Do you wish to proceed?[n]:
```

CLEinstall may resolve some issues after you indicate that you want to proceed; for example, disk devices are already mounted, boot image file or links already exist, HSS daemons are stopped on the SMW.



CAUTION: Some problems can be resolved only through manual intervention via another terminal window or by rebooting the SMW; for example, a process is using a mounted disk partition, preventing CLEinstall from unmounting the partition.

5. Monitor the debug output. Create another terminal window and invoke the `tail` command by using the path and timestamp displayed when CLEinstall was run.

```
smw~:# tail -f /var/adm/cray/logs/CLEinstall.p#. YYYYMMDDhhmmss. $LABEL.debug.log
```

6. Locate the following warning and prompt in the CLEinstall console window and type `y`.

```
*** Preparing to UPGRADE software on system set label system_set_label. Do you wish to proceed? [n]
```

The CLEinstall program now installs the release software. This command runs for 30 minutes or more for updates and 90 minutes for an upgrade, depending on your system configuration.

7. Monitor the output to ensure that your installation is proceeding without error. Several error messages from the `tar` command are displayed as the persistent `/var` is updated for each service node. You may safely ignore these messages.
8. Confirm that the CLEinstall program has completed successfully.

On completion, the CLEinstall program generates a list of command hints to be run as the next steps in the update or upgrade process. These commands are customized, based on the variables in the CLEinstall.conf and sysset.conf files, and include runtime variables such as PID numbers in file names. The list of command hints is written to the CLEinstall.command_hints.timestamp file in the installer log directory.

Complete the upgrade/update and configuration of your Cray system by using both the commands that the CLEinstall program provides and the information in the remaining sections of this chapter.

As you complete these procedures, you can cut and paste the suggested commands from the output window or from the window created in a previous step that tailed the debug file. The log files created in `/var/adm/cray/logs` for CLEinstall.P#. YYYYMMDDhhmmss. \$LABEL.stdout.log and CLEinstall.P#. YYYYMMDDhhmmss. \$LABEL.debug.log also contain the suggested commands.

Create Boot Images

The Cray CNL compute nodes and Cray service nodes use RAM disks for booting. Service nodes and CNL compute nodes use the same `initramfs` format and workspace environment. This space is created in `/opt/xt-images/machine-xtrelease-LABEL-partition/nodetype`, where `machine` is the Cray hostname, `xtrelease` is the build level for the CLE release, `LABEL` is the system set label used from `/etc/sysset.conf`, `partition` describes either the full machine or a system partition, and `nodetype` is either `compute` or `service`.



CAUTION: Existing files in `/opt/xt-images/templates/default` are copied into the new bootimage work space. In most cases, you can use the older version of the files with the upgraded system. However, some file content may have changed with the new release. Verify that site-specific modifications are compatible. For example, use existing copies of `/etc/hosts`, `/etc/passwd` and `/etc/modprobe.conf`, but if `/init` changed for the template, the site-modified version that is copied and used for CLE 5.2 may cause a boot failure.

Follow the procedures in this section to prepare the work space in `/opt/xt-images`. For more information about configuring boot images for service and compute nodes, see the `xtclone(8)` and `xtpackage(8)` man pages.

Prepare Compute and Service Node Boot Images

The `shell_bootimage_LABEL.sh` script prepares boot images for the system set specified by `LABEL`. For example, if your system set has the label `BLUE` in `/etc/sysset.conf`, invoke `shell_bootimage_BLUE.sh` to prepare a boot image. This script uses `xtclone` and `xtpackage` to prepare the work space in `/opt/xt-images`.

NOTE: When upgrading a system using direct-attached Lustre (DAL), use the `-d` option. This option specifies that the CentOS DAL be included.

For more information about configuring boot images for service and compute nodes, see the `xtclone(8)` and `xtpackage(8)` man pages.

1. Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

2. Run the `shell_bootimage_*.sh` script, where `LABEL` is the system set label specified in `/etc/sysset.conf` for this boot image.

Specify the `-c` option to automatically create and set the boot image for the next boot. For example, if the system set label is `BLUE`:

```
smw:~# /var/opt/cray/install/shell_bootimage_BLUE.sh -c
```

For information about additional options accepted by this script, use the `-h` option to display a help message.

Enable Boot Node Failover

NOTE: Boot node failover is an optional CLE feature.

If boot-node failover has been configured for the first time, follow these steps. If boot-node failover has not been configured, skip this procedure.

To enable bootnode failover, you must set `bootnode_failover` parameters in the `CLEinstall.conf` file before you run the `CLEinstall` program. For more information, see [Configure Boot Node Failover](#) on page 81.

In this example, the primary boot node is `c0-0c0s0n1` (`node_boot_primary=1`) and the backup or alternate boot node is `c0-0c1s1n1` (`node_boot_alternate=61`).

TIP: Use the `rtr --system-map` command to translate between NIDs and physical ID names.

1. As `crayadm` on the SMW, halt the primary and alternate boot nodes.



WARNING: Verify that the system is shut down before you invoke the `xtcli halt` command.

```
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
```

- Specify the primary and backup boot nodes in the boot configuration.

If the partition variable in CLEinstall.conf is s0, type the following command to select the boot node for the entire system.

```
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

Or

If the partition variable in CLEinstall.conf is a partition value such as p0, p1, and so on, type the following command to select the boot node for the designated partition.

```
crayadm@smw:~> xtcli part_cfg update pN -b c0-0c0s0n1,c0-0c1s1n1
```

- To use boot-node failover, enable the STONITH capability on the blade or module of the primary boot node. Use the xtdaemonconfig command to determine the current STONITH setting.

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke xtdaemonconfig with the --partition pn option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary boot node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

- To enable STONITH on the primary boot node blade, type the following command:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 stonith=true
c0-0c0s0: stonith=true
crayadm@smw:~> xtcli halt c0-0c0s0n1,c0-0c1s1n1
crayadm@smw:~> xtcli boot_cfg update -b c0-0c0s0n1,c0-0c1s1n1
```

NOTE: If the system is partitioned, invoke xtdaemonconfig with the --partition pn option.

Enable SDB Node Failover

NOTE: SDB node failover is an optional CLE feature.

If SDB node failover has been configured for the first time, follow these steps. If SDB node failover has not been configured, skip this procedure.

In addition to this procedure, refer to [Configure Boot Automation for SDB Node Failover](#) on page 140 after you have completed the remaining configuration steps and have booted and tested your system.

To enable SDB node failover, you must set sdbnode_failover parameters in the CLEinstall.conf file before you run the CLEinstall program. For more information, see [Configure SDB Node Failover](#) on page 82.

In this example, the primary SDB node is c0-0c0s2n1 (node_sdb_primary=5) and the backup or alternate SDB node is c0-0c1s3n1 (node_sdb_alternate=57).

TIP: Use the rtr --system-map command to translate between NIDs and physical ID names.

- Invoke xtdaemonconfig to determine the current STONITH setting on the blade or module of the primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s0 | grep stonith
c0-0c0s0: stonith=false
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.



CAUTION: STONITH is a per blade setting, not a per node setting. You must ensure that your primary SDB node is not assigned to a blade that hosts services with conflicting STONITH requirements, such as Lustre.

2. Enable STONITH on your primary SDB node. For example:

```
crayadm@smw:~> xtdaemonconfig c0-0c0s2 stonith=true
c0-0c0s2: stonith=true
The expected response was received.
```

NOTE: If the system is partitioned, invoke `xtdaemonconfig` with the `--partition pn` option.

3. Specify the primary and backup SDB nodes in the boot configuration.

For example, if the partition variable in `CLEinstall.conf` is `s0`, type the following command to select the primary and backup SDB nodes.

```
crayadm@smw:~> xtcli halt c0-0c0s2n1,c0-0c1s3n1
crayadm@smw:~> xtcli boot_cfg update -d c0-0c0s2n1,c0-0c1s3n1
```

Or

If the partition variable in `CLEinstall.conf` is a partition value such as `p0`, `p1`, and so on, type the following command:

```
crayadm@smw:~> xtcli part_cfg update pN -d c0-0c0s2n1,c0-0c1s3n1
```

Update Direct-Attached Lustre

NOTE: The direct-attached Lustre (DAL) file system is optional; your storage RAID may be external to the mainframe.

Installation and configuration of DAL differs from that of external Lustre. Cray service nodes that support DAL use a CentOS operating system running on ramdisk as opposed to the shared root file system.

1. Build the DAL image root.

```
smw:~ # impscli build image_recipe dal_cle_5.2up04_centos_6.5_x86-64_ari
```

2. Update the config set. The IMPS Configurator updates the config set by interactively guiding the administrator through the process of providing needed configuration values. As the Configurator prompts for information about the system, a description and guidance, including a reasonable default value, are provided for each query.

```
smw:~ # impscli update config_set pN with images \
dal_cle_5.2up04_centos_6.5_x86-64_ari
```

Where `pN` is a valid partition name.

3. Provision the DAL image.

```
smw:~ # impscli provisiondal image dal_cle_5.2up04_centos_6.5_x86-64_ari to \
/opt/xt-images/machine-xtrelease-LABEL-partition
```

Where *LABEL* is the same as in [Prepare Compute and Service Node Boot Images](#) on page 374.

Informational messages are displayed. Warning messages, related to creating a CentOS image on an SLES system, are also displayed and can safely be ignored. Finally, a message similar to the following is displayed after the provisioning completes successfully.

```
INFO - Provisioning of DAL image 'dal_cle_5.2up04_centos_6.5_x86-64_ari'
successful.
```

4. Recreate the boot image to include DAL.

```
smw:~ # /var/opt/cray/install/shell_bootimage_LABEL.sh -d -c -b /bootimagedir/
bootimage.cpio
```

Where *LABEL* is the same as in [Prepare Compute and Service Node Boot Images](#) on page 374.

5. Copy the boot package on the SMW to the same directory on the boot node.

```
smw:~ # cp -p /bootimagedir/bootimage.cpio /bootrootdir/bootimagedir/bootimage.cpio
```

6. Remember to update the version level of the boot image for the DAL service nodes when editing the boot automation file `/opt/cray/hss/default/etc/auto.xthostname` during [Boot and Test the System](#) on page 378.

```
lappend actions [list crms_boot_loadfile \
dal_cle_5.2up04_centos_6.5_x86-64_ari service \
cnames_of_your_DAL_nodes linux]
```

Where *cnames_of_your_DAL_nodes* is a comma-separated list.

7. Unmount the CentOS media.

```
smw:~ # umount /media/Centosbase
```

Run Post-CLEinstall Commands

1. Unmount and eject the release software DVD from the SMW DVD drive if it is still loaded.

```
smw:~# umount /media/cdrom
smw:~# umount /media/Centosbase
smw:~# eject
```

2. Run the `shell_post_install.sh` script on the SMW to unmount the boot root and shared root file systems and perform other cleanup as needed.

```
smw:~# /var/opt/cray/install/shell_post_install.sh /bootroot0 /sharedroot0
```



WARNING: Exercise care when you mount and unmount file systems. If you mount a file system on the SMW and boot node simultaneously, you may corrupt the file system.

3. Confirm that the `shell_post_install.sh` script successfully unmounted the boot root and shared root file systems.

If a file system does not unmount successfully, the script displays information about open files and associated processes (by using the `lsof` and `fuser` commands). Attempt to terminate processes with open files and if necessary, reboot the SMW to resolve the problem.

Configure Optional Services

If you enabled an optional service you were not previously using in [Prepare the CLEinstall.conf Configuration File](#) on page 296, you may need to perform additional configuration steps. Follow the procedures in the appropriate optional section in [Install CLE on a New System](#) on page 91 or in [Managing System Software for the Cray Linux Environment \(S-2393\)](#).

If you configured an optional CLE feature or service during a previous installation or upgrade, no additional steps are required.

Configure MAMU Nodes

1. On the boot node, run this script, which ensures the keys are correct on the MAMU nodes (the `postproc` node class in this and previous examples).

```
boot# /var/opt/cray/install/shell_ssh.sh
```

2. Modify the `sshd_config` and `/etc/fstab` files for the new `postproc` class.

```
boot# xtopview -c postproc -m "setting up postproc nodes"
class/postproc:/# xtspec /etc/fstab
```

Add this line:

```
ufs: /ufs/home          /ufs/home      nfs           tcp,rw      0 0
```

```
class/postproc:/ # xtspec /etc/ssh/sshd_config
class/postproc:/ # vi /etc/ssh/sshd_config
```

Strip any `MatchUser` blocks from the bottom of the `sshd_config` file. Save and close the file.

3. Run these commands to restrict logins on the `postproc` nodes to only the `crayadm` administrative account and `root`, which is necessary to provide out of memory protection.

```
class/postproc:/ # xtspec /etc/ssh/sshd_config
class/postproc:/ # echo "AllowUsers root crayadm" >> /etc/ssh/sshd_config
class/postproc:/ # exit
```

Boot and Test the System

IMPORTANT: If you configured optional services for the first time during this update or upgrade and deferred updating the boot image, update the boot image now by following [Prepare Compute and Service Node Boot Images](#) on page 374.

Your system is now upgraded.

Reboot the Cray System

1. Use site-specific procedures to shut down the system. For example, to shutdown using an automation file type the following:

```
crayadm@smw:~> xtbootsys -s last -a auto.xtshutdown
```

For more information about using automation files see the `xtbootsys(8)` man page.

Although not the preferred method, alternatively execute these commands as `root` from the boot node to shutdown your system.

```
boot:~ # xtshutdown -y
boot:~ # shutdown -h now;exit
```

2. Edit the boot automation file and make site-specific changes as needed.

```
crayadm@smw:~> vi /opt/cray/hss/default/etc/auto.xhostname
```

3. Use the `xtbootsys` command to boot the Cray system.



CAUTION: Shut down your Cray system before invoking the `xtbootsys` command. If installing to an alternate system set, shut down the currently running system before booting the new boot image.

Type this command to boot the entire system.

```
crayadm@smw:~> xtbootsys -a auto.xhostname
```

Or

Type this command to boot a partition.

```
crayadm@smw:~> xtbootsys --partition pN -a auto.xhostname
```

Flash the nvBIOS for Kepler GPUs

A Cray XC30 system with NVIDIA® Tesla® SXM modules requires an update to the NVIDIA BIOS (nvBIOS) for the NVIDIA K20X and K40s graphics processing units (GPUs). The nvBIOS is unique for each SXM-1 Kepler™ SKU, based on the type of heat sink, as shown below.

GPU Type	Board SKU	Production Firmware Image Version
Kepler K20X (13 fin)	P2085 SKU 202	80.10.44.00.02
Kepler K20X (20 fin)	P2085 SKU 212	80.10.44.00.04
Kepler K20X (30 fin)	P2085 SKU 222	80.10.44.00.05
Kepler K40s (13 fin)	P2085 SKU 209	80.80.4B.00.03
Kepler K40s (20 fin)	P2085 SKU 219	80.80.4B.00.04
Kepler K40s (30 fin)	P2085 SKU 229	80.80.4B.00.05

The CLE software includes a script that automatically determines the SKU version and flashes the nvBIOS with the appropriate firmware.

TIP: You can use the `cselect` command to identify the number and location of the Kepler GPUs. This example shows a system with K20X GPUs on four nodes.

```
login:~# cselect -c "subtype.eq.'nVidia_Kepler'"
4
```

```
login:~# cselect -e "subtype.eq.'nVidia_Kepler'"
70-73
```

1. As root on the login node, set the allocation mode for all compute nodes to interactive.

```
login:~# xtprocadmin -km interactive
```

2. Change to the directory containing the NVIDIA scripts.

```
login:~# cd /opt/cray/cray-nvidia/default/bin
```

3. To flash the Kepler K20X GPUs, for example, choose one of the following options.

- To update the entire system:

```
login:~# aprun -n `cselect -c "subtype.eq.'nVidia_Kepler'"` \
-N 1 -L `cselect -e "subtype.eq.'nVidia_Kepler'"` \
./nvFlashBySKU -b
```

- To update a single node or set of nodes:

```
login:~# aprun -n PEs -N 1 -L node_list ./nvFlashBySKU -b
```

NOTE: Replace *PEs* with the number of nodes; replace *node_list* with a comma-separated list of NIDs.

For example, to flash four GPUs on nodes 70-73:

```
login:~# aprun -n 4 -N 1 -L 70,71,72,73 ./nvFlashBySKU -b
c0-0c0s1n0: Nid 70: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n1: Nid 71: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n2: Nid 72: Successful Cray Graphite K20X nvBIOS flash
c0-0c0s1n3: Nid 73: Successful Cray Graphite K20X nvBIOS flash
```

4. If there is a flash failure, nvFlashBySKU displays an error message with the failing node ID, as in this example:

```
c0-0c0s1n3: Nid 73: Failed Cray Graphite K20X nvBIOS flash
```

Depending on the type of failure, nvFlashBySKU might display additional information, if available. No flashing is done on unsupported SKUs.

If a GPU fails to flash, the SXM-1 card must be replaced.

5. After flashing is successful, use xtbootstys to reboot the nodes from the SMW. For example:

```
crayadm@smw:~> xtbootstys --reboot -L CNL0 -r "rebooting after nvBIOS update" \
c0-0c0s1n0,c0-0c0s1n1,c0-0c0s1n2,c0-0c0s1n3
```

TIP: You can use xtprocadmin on the login node to determine each node name from the cselect output, as in this example:

```
login:~# xtprocadmin -n `cselect -e "subtype.eq.'nVidia_Kepler'"`
  NID      (HEX)      NODENAME      TYPE      STATUS      MODE
   70      0xf8      c0-0c0s1n0    compute    up          batch
   71      0xf9      c0-0c0s1n1    compute    up          batch
   72      0xfa      c0-0c0s1n2    compute    up          batch
   73      0xfb      c0-0c0s1n3    compute    up          batch
```



```

c1n0 ;;;;;;;;; ;;;;;;;;; ;S;S;S;S ;;;;;;;;;
n3 S;S;S;S; ;;;;;;;;; ;;;;;;;;; ;;;;;;;;;
n2 S;S;S;S; ;;;;;;;;; ;;;;;;;;; ;;;;;;;;;
n1 S;S;S;S; ;;;;;;;;; ;;;;;;;;; ;;;;;;;;;
c0n0 S;S;S;S; ;;;;;;;;; ;;;;;;;;; ;;;;;;;;;
s01234567 01234567 01234567 01234567

```

Legend:

```

nonexistent node          S  service node
; free interactive compute node  - free batch compute node
A allocated, but idle compute node ? suspect compute node
X down compute node       Y  down or admindown service node
Z admindown compute node

```

Available compute nodes: 352 interactive, 0 batch

The output for Cray XC30 systems follows.

```

crayadm@login:~> xtnodestat
Current Allocation Status at Fri Feb 15 15:01:38 2013

      C0-0          C1-0          C2-0          C3-0
n3  ---a-----a---  ---a-----a---  --aa-----a---  -----
n2  ---a-----a---  ---a-----a---  --aa-----a---  -----
n1  ---a-----a---  ---a-----a---  --Xa-----a---  -----
c2n0 ---a-----a---  ---a-----a---  --aa-----a---  -----
n3  ----bX--a-a---  ---a-----a-a-  - a-----a---  ---a-----
n2  ----ba--a-a---  ---a-----a-a-  -Sa-----a---  ---X-----
n1  ----ba--a-a---  ---a-----a-a-  -Sa-----a---  ---a-----
c1n0 ----ba--a-a---  ---a-----X-a-  - a-----a---  ---a-----
n3  -a-----a---  ---a-----a---  - a-----a---  --a-----a---
n2  SS-a-----a---  SS-----a---  -S-a-----  S--X-----a---
n1  SS-X-----a---  SS-----a---  -S-a-----  S--a-----Aa---
c0n0 -a-----a---  ---a-----a---  - -X-----  --a-----Aa---
s0123456789abcdef 0123456789abcdef 0123456789abcdef 0123456789abcdef

```

Legend:

```

nonexistent node          S  service node
; free interactive compute node  - free batch compute node
A allocated (idle) compute or ccm node ? suspect compute node
W waiting or non-running job      X  down compute node
Y down or admindown service node  Z  admindown compute node

```

Available compute nodes: 0 interactive, 650 batch

The xtprocadmin command displays the current values of processor flags and node attributes. The output for Cray XE and Cray XK systems follows.

```

crayadm@login:~> xtprocadmin
  NID   (HEX)  NODENAME    TYPE    STATUS    MODE
    0     0x0   c0-0c0s0n0  service  up       interactive
    2     0x2   c0-0c0s1n0  service  up       interactive
    4     0x4   c0-0c0s2n0  service  up       interactive
    6     0x6   c0-0c0s3n0  service  up       interactive
. . .
   93    0x5d   c0-0c2s1n3  service  up       interactive
   94    0x5e   c0-0c2s0n2  service  up       interactive
   95    0x5f   c0-0c2s0n3  service  up       interactive

```

The output for Cray XC30 systems follows.

```
crayadm@login:~> xtprocadmin
  NID      (HEX)      NODENAME      TYPE      STATUS      MODE
  1         0x1      c0-0c0s0n1  service   up          batch
  2         0x2      c0-0c0s0n2  service   up          batch
  5         0x5      c0-0c0s1n1  service   up          batch
  6         0x6      c0-0c0s1n2  service   up          batch
  8         0x8      c0-0c0s2n0  compute   up          batch
  9         0x9      c0-0c0s2n1  compute   up          batch
  10        0xa      c0-0c0s2n2  compute   up          batch
```

The `apstat` command displays the current status of all applications running on the system.

```
crayadm@login:~> apstat -v
Compute node summary
  arch config  up  resv  use  avail  down
  XT    733    733  107   89   626    0

Total pending applications: 4
Pending Pid      User      w:d:N NID      Age Command  Why
  17278  crayadm  1848:1:24  5    0h53m  ./app1  Busy
  17340  crayadm  1848:1:24  5    0h53m  ./app1  Busy
  17469  crayadm  1848:1:24  5    0h52m  ./app1  Busy
  26155  crayadm  1848:1:24  5    0h12m  ./app2  Busy

Total placed applications: 2
  Apid ResId  User  PEs Nodes  Age State  Command
  1631095  135 alan-1  64  4  0h31m  run  mcp
  1631145  140 flynn  128  8  0h05m  run  TRON-JA307020
```

4. Run a simple job on the compute nodes.

At the conclusion of the installation process, the `CLEinstall` program provides suggestions for runtime commands and indicates how many compute nodes are available for use with the `aprun -n` option.

For `aprun` to work cleanly, the current working directory on the login node should also exist on the compute node. Change your current working directory to either `/tmp` or to a directory on a mounted Lustre file system.

For example, type the following.

```
crayadm@login:~> cd /tmp
crayadm@login:~> aprun -b -n 16 -N 1 /bin/cat /proc/sys/kernel/hostname
```

This command returns the hostname of each of the 16 compute nodes used to execute the program.

```
nid00010
nid00011
nid00012
nid00020
nid00016
nid00040
nid00052
nid00078
nid00084
nid00043
nid00046
nid00049
. . .
```

5. Test file system functionality. For example, if you have a Lustre file system named `/mylusmnt/filesystem`, type the following.

```

crayadm@login:~> cd /mylustmnt/filesystem
crayadm@login:/mylustremnt/filesystem> echo lustretest > testfile
crayadm@login:/mylustremnt/filesystem> aprun -b -n 5 -N 1 /bin/cat ./testfile
lustretest
lustretest
lustretest
lustretest
lustretest
Application 109 resources: utime ~0s, stime ~0s

```

6. Test the optional features that you have configured on your system.

- a. To test RSIP functionality, log on to an RSIP client node (compute node) and ping the IP address of the SMW or other host external to the Cray system. For example, if c0-0c0s7n2 is an RSIP client, type the following commands.

```

crayadm@login:~> exit
boot:~ # ssh root@c0-0c0s7n2
root@c0-0c0s7n2's password:
Welcome to the initramfs
# ping 172.30.14.55
172.30.14.55 is alive!
# exit
Connection to c0-0c0s7n2 closed.
boot:~ # exit

```

NOTE: RSIP clients on the compute nodes make connections to the RSIP server(s) during system boot. Initiation of these connections is staggered over a two minute window; during that time, connectivity over RSIP tunnels is unreliable. Avoid using RSIP services for three to four minutes following a system boot.

- b. To check the status of DVS, type the following command on the DVS server node.

```

crayadm@login:~> ssh root@nid00019 /etc/init.d/dvs status
DVS service: ..running

```

To test DVS functionality, invoke the mount command on any compute node.

```

crayadm@login:~> ssh root@c0-0c0s7n2 mount | grep dvs
/dvs-shared on /dvs type dvs
(rw,blksize=16384,nodename=c0-0c0s4n3,nocache,nodatasync,\
retry,userenv,clusterfs,maxnodes=1,nnodes=1)

```

Create a test file on the DVS mounted file system. For example, type the following.

```

crayadm@login:~> cd /dvs
crayadm@login:/dvs> echo dvstest > testfile
crayadm@login:/dvs> aprun -b -n 5 -N 1 /bin/cat ./testfile
dvstest
dvstest
dvstest
dvstest
dvstest
Application 121 resources: utime ~0s, stime ~0s

```

- Following a successful installation, the file `/etc/opt/cray/release/clerelease` is populated with the installed release level. For example,

```
crayadm@login:~> cat /etc/opt/cray/release/clerelease
5.2.UP04
```

If the preceding simple tests ran successfully, the system is operational. Cray recommends using the `xthotbackup` utility to create a backup of a newly updated or upgraded system. For more information, see the `xthotbackup(8)` man page.

Copy Boot Images and CLE Install Directory to the Passive SMW

- If boot images are stored as files, log on to `smw1` as `root` and copy the boot image to the other SMW. This manual copy operation speeds up future synchronization.

NOTE: In this command, replace `smw1` with the host name of the active SMW, and replace `smw2` with the host name of the passive SMW. Replace `bootimagedir` with the name of the boot image directory, and replace `file` with the name of the boot image.

```
smw1:~ # scp -p /bootimagedir/file smw2:/bootimagedir/file
```

IMPORTANT: The `bootimagedir` directory must already exist on the passive SMW.

- Copy the CLE install directory, `/home/crayadm/install.xtre1`, from the first SMW to a local directory on the second SMW (such as `/tmp`). Do not use `/home/crayadm` on the second SMW, because that would create local differences for this shared directory. Replace `xtre1` with the site-determined name specific to the release being installed.

Upgrade CLE Software on the Passive SMW

Use the following procedures to upgrade the CLE software on the passive SMW.

Install CLE Release Software on the SMW

Two DVDs are provided to install the CLE 5.2 release on a Cray system. The first is labeled `Cray CLE 5.2.UPnn Software` and contains software specific to Cray systems. Optionally, you may have an ISO image called `xc-sles11sp3-5.2.55d05.iso`, where `5.2.55` indicates the CLE release build level, and `d05` indicates the installer version.

The second DVD is labeled `CentOS-6.5-x86_64-bin-DVD1.iso` and contains the CentOS 6.5 base operating system for CLE direct-attached Lustre (DAL) nodes.

Copy the Software to the SMW

- Log on to the SMW as `root`.

```
crayadm@smw:~> su - root
```

- Mount the release media by using one of the following commands, depending on your media type.

If installing the release package from disk, place the Cray CLE 5.2.UPnn Software DVD in the CD/DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the release media using the ISO image, execute the following command, where `xc-sles11sp3-5.2.55d05.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro xc-sles11sp3-5.2.55d05.iso /media/cdrom
```

3. Copy all files to a directory on the SMW in `/home/crayadm/install.xtrel`, where `xtrel` is a site-determined name specific to the release being installed. For example:

```
smw:~# mkdir /home/crayadm/install.5.2.55
smw:~# cp -pr /media/cdrom/* /home/crayadm/install.5.2.55
```

4. Unmount the Cray CLE 5.2.UPnn Software media.

```
smw:~# umount /media/cdrom
```

5. For upgrading from CLE 5.1 or CLE 4.2 to CLE 5.2, you must mount the SLES 11 SP3 base media. Insert the Cray-CLEbase11sp3 DVD into the SMW DVD drive and mount it.

```
smw:~# mount /dev/cdrom /media/cdrom
```

Or

To mount the base operating system media using the ISO image, execute the following command, where `Cray-CLEbase11SP3-yyyymmdd.iso` is the path name to the ISO image file.

```
smw:~# mount -o loop,ro Cray-CLEbase11SP3-yyyymmdd.iso /media/cdrom
```

6. For direct-attached Lustre implementations, also mount the `CentOS-6.5-x86_64-bin-DVD1.iso` image.

```
smw:~# mkdir /media/Centosbase
smw:~# mount -o loop,ro CentOS-6.5-x86_64-bin-DVD1.iso /media/Centosbase
```

Install CLE on the SMW

1. As `root`, execute the `CRAYCLEinstall.sh` installation script to upgrade the Cray CLE software on the SMW.

```
smw:~# /home/crayadm/install.5.2.56/CRAYCLEinstall.sh \
-m /home/crayadm/install.5.2.56 -u -v -w
```

2. At the prompt `Do you wish to continue?`, type `y` and press `Enter`.

The output of the installation script is displayed to the console. If this script fails, you can restart it with the same options. However, rerunning this script may generate numerous error messages as it attempts to install already-installed RPMs. You may safely ignore these messages.

Finish the CLE Upgrade

After upgrading the CLE software on the passive SMW, use this procedure to finish the upgrade on the SMW HA system.

Add the following files to the synchronization list:

```
/var/opt/cray/install/shell_bootimage*  
/var/opt/cray/install/networking_configuration-p*.json
```

For this procedure, see [Add Site-specific Files to the Synchronization List](#).

Configure PMDB Storage

Choose one of these options to configure shared storage for the Power Management Database (PMDB).

- [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172. Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.
- [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177. Shared storage: A logical disk, configured as a LUN (Logical Unit) or logical volume on the boot RAID. The boot RAID must have sufficient space for `/var/lib/pgsql`.

Cray strongly recommends using either mirrored storage (preferred) or shared storage. An unshared PMDB is split across both SMWs; data collected before an SMW failover will be lost or not easily accessible after failover. For more information, see [Storage for the Power Management Database \(PMDB\)](#) on page 9.

Configure Mirrored Storage with DRBD for the PMDB

Prerequisites

IMPORTANT:

If mirrored storage becomes available after the PMDB has been configured for shared storage, use the procedure [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322 instead of this procedure.

Before beginning this procedure:

- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- Plan sufficient time for this procedure. Transferring the Power Management Database (PMDB) to a 1 TB disk requires about 10 hours. The SMW HA cluster should be in maintenance mode until the synchronization operation completes. The Cray system (compute and service nodes) can remain up and can run jobs during this period.
- Check `/etc/fstab` to ensure that there is no entry for `phy3`.

- If upgrading or updating the SMW HA system, ensure that the following RPMs are installed on both SMWs and that the version number is 8.4.4 or higher:

```
drbd-bash-completion-8.4.4-0.22.9
drbd-kmp-default-8.4.4_3.0.101_0.15-0.22.7
drbd-udev-8.4.4-0.22.9
drbd-utils-8.4.4-0.22.9
drbd-pacemaker-8.4.4-0.22.9
drbd-xen-8.4.4-0.22.9
drbd-8.4.4-0.22.9
```

If necessary, install or update any missing RPMs with "zypper install drbd".

Mirrored storage (preferred): An optional pair of disks, one in each SMW, to store PMDB data. In this configuration, the active SMW mounts `/var/lib/pgsql` as a Distributed Replicated Block Device (DRBD) device and communicates replicated writes over a private TCP/IP connection (`eth5`) to the passive SMW. This is the preferred PMDB configuration to ensure availability of the PMDB data without competition for I/O bandwidth to the SMW root disk or boot RAID file systems.

This procedure configures the network for DRBD, configures the DRBD disks, and transfers the PMDB data from local disk to the mirrored DRBD disks.

1. Add `eth5` to the network files.

- a. Log in as root on the first SMW (`smw1`).

```
workstation> ssh root@smw1
```

- b. On `smw1`, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.2/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

- c. In a separate terminal session, log in as root on the other SMW (`smw2`).

```
workstation> ssh root@smw2
```

- d. On `smw2`, create the file `/etc/sysconfig/network/ifcfg-eth5` and add the following contents.

```
BOOTPROTO='static'
IPADDR='10.5.1.3/16'
NAME='eth5 SMW HA DRBD Network'
PREFIXLEN='16'
STARTMODE='auto'
USERCONTROL='no'
```

2. Reinitialize the `eth5` interface on both SMWs.

```
smw1:~# ifdown eth5; sleep 1; ifup eth5
```

```
smw2:~# ifdown eth5; sleep 1; ifup eth5
```

3. Verify the IP addresses from *smw1*.

```
smw1:~# ping -c3 10.5.1.3
```

4. Configure the firewall to allow eth5 as an internal connection on both SMWs.

- a. Edit the file `/etc/sysconfig/SuSEfirewall12` on both *smw1* and *smw2*.
- b. Locate the line containing the `FW_DEV_INT` variable.
- c. If necessary, add `eth5` to the end of the `FW_DEV_INT` line.

```
FW_DEV_INT="eth1 eth2 eth3 eth4 eth5 lo"
```

- d. Save your changes and exit the editor on both SMWs.

5. Reinitialize the IP tables by executing the `/sbin/SuSEfirewall12` command on both SMWs.

```
smw1:~# /sbin/SuSEfirewall12
```

```
smw2:~# /sbin/SuSEfirewall12
```

6. On the active SMW only, add the new DRDB disk to the SMW HA configuration.

NOTE: The following examples assume that *smw1* is the active SMW.

- a. Verify that the device exists on both SMWs.
For Dell R-630 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1
ls -l /dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0
```

For Dell R815 systems:

```
smw1:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

```
smw2:~# ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
ls -l /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

- b. Determine if the dedicated disk for the PMDB must be formatted. In this procedure, this disk is referred to as `PMDISK`.

NOTE: If the `PMDISK` is already correctly formatted, skip to step [6.f](#) on page 390.

This procedure assumes that a disk drive is available for use as a dedicated drive for the PMDB. The drive should be physically located within the rack-mount SMW at slot 4. The drive should be of the specification 1 TB 7.2K RPM SATA 3Gbps 2.5in HotPlug Hard Drive 342-1998, per the SMW Bill of Materials. On a Dell PowerEdge R815 the device for `PMDISK`

is `/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0` On a Dell PowerEdge R630 the device for `PMDISK` is `/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0`.

- c. Verify that the `PMDISK` is inserted into the SMW by entering the correct device name. This example is for a Dell R815.

```
smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
```

```

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
255 heads, 63 sectors/track, 121601 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1

```

Device	Boot	Start	End	Blocks	Id	System
--------	------	-------	-----	--------	----	--------

- d. Create a new primary partition for the PMDISK, and write it to the partition table. If there are any existing partitions on this disk, manually delete them first.

```

smw:#fdisk \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0
Command (m for help): n
Command action
  e   extended
  p   primary partition (1-4)p
Partition number (1-4, default 1): 1
First sector (2048-1953525167, default 2048): [press return]
Using default value 2048
Last sector, +sectors or +size{K,M,G} (2048-1953525167, default 1953525167): [press
return]
Using default value 1953525167
Command (m for help): w
The partition table has been altered!
Calling ioctl() to re-read partition table.
Syncing disks.

```

- e. Verify that the partition has been created. This should be device /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1

```

smw:#fdisk -l \
/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0

Disk /dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0:
1000.2 GB, 1000204886016 bytes
81 heads, 63 sectors/track, 382818 cylinders, total 1953525168 sectors
Units = sectors of 1 * 512 = 512 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0xffdfd1e1

                Device Boot      Start         End      Blocks   Id  System
/dev/disk/by-path/. . .-lun-0-part1    2048    1953525167    976761560   83  Linux

```

- f. Navigate to the directory containing the SMWHAconfig command.

```
smw1:~# cd /opt/cray/ha-smw/default/hainst
```

- g. Execute SMWHAconfig to add the DRBD disk. For *disk-device*, specify the disk ID of the disk backing the DRBD disk, using either the by-name or by-path format for the device name.

On a rack-mount SMW (either Dell R815 or R630), the DRBD disk is a partition on the disk in slot 4. On a Dell 815 this is `/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1`. On a Dell 630 it is `/dev/disk/by-path/pci-0000:03:00.0-scsi-0:0:4:0-part1`

```
smw1:~# ./SMWHAconfig --add_disk=pm-fs --device=/dev/drbd_r0 --directory=/var/lib/pgsql \
--pm_disk_name=/dev/disk/by-path/pci-0000:05:00.0-sas-phy3-0x4433221103000000-lun-0-part1
```

7. Reboot the active SMW (*smw1*) and wait for it to boot completely.
8. Reboot the other SMW (*smw2*) and wait for it to boot completely.
9. Correct the permissions for the `/var/lib/pgsql` file on the active SMW.

```
smw1:~# chown postgres:postgres /var/lib/pgsql
smw1:~# chmod 750 /var/lib/pgsql
```

10. Put the SMW HA cluster into maintenance mode while waiting for the DRBD sync operation to complete. When *smw1* and *smw2* rejoin the cluster after rebooting, the primary DRBD disk (in *smw1*) synchronizes data to the secondary disk (in *smw2*). DRBD operates at the device level to synchronize the entire contents of the PMDB disk. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. The Cray system (service and compute nodes) can be booted and can run jobs during this period.

IMPORTANT:

Cray strongly recommends putting the SMW HA cluster into maintenance mode to prevent any failover during the sync operation. If a failover were to occur during this period, the newly-active SMW could have an incomplete copy of PMDB data.

- a. Put the SMW HA cluster into maintenance mode on *smw1*.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- b. Check the status of the DRBD sync operation with either `rcdrbd status` or `cat /proc/drbd`. The `rcdrbd` output is easier to read, but `/proc/drbd` contains more status information and includes an estimate of time to completion.

```
smw1:~# rcdrbd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res  cs          ro          ds          p
mounted          fstype
0:r0   SyncSource   Primary/Secondary  UpToDate/Inconsistent  C  /var/lib/
pgsql  ext3
...   sync'ed:    72.7%          (252512/922140)M
```

```
smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2
ua:0 ap:0 ep:1 wo:f oos:260636656
[=====>.....] sync'ed: 72.4% (254524/922140)M
finish: 2:21:07 speed: 30,768 (29,720) K/sec
```

For an explanation of the status information in `/proc/drbd`, see the DRDB User's Guide at [linbit.com: http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd](http://drbd.linbit.com/users-guide/ch-admin.html#s-proc-drbd).

- When the DRBD sync operation finishes, bring the HA cluster out of maintenance mode on `smw1`.

```
smw1:~# crm configure property maintenance-mode=false 2> /dev/null
```

- Examine the output of `crm status` to ensure that the `ip_drbd_pgsql` is started on `smw1` and that the `Masters` and `Slaves` entries for `ms_drbd_pgsql` display the SMW host names (`smw1` and `smw2`).

```
smw1:~# crm status
Last updated: Thu Jan 22 18:40:21 2015
Last change: Thu Jan 22 11:51:36 2015 by hacluster via crmd on smw1
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.11-3ca8c3b
2 Nodes configured, 2 expected votes
23 Resources configured

Online: [ smw1 smw2 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
.
.
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (lsb:cray-syslog):             Started smw1
  homedir     (ocf::heartbeat:Filesystem):    Started smw1
  md-fs       (ocf::heartbeat:Filesystem):    Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):    Started smw1
  postgresql  (lsb:postgresql):             Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1
ip_drbd_pgsql (ocf::heartbeat:IPaddr2):      Started smw1
Master/Slave Set: ms_drbd_pgsql [drbd_pgsql]
  Masters: [ smw1 ]
  Slaves: [ smw2 ]
```

Configure Shared Storage on the Boot RAID for the PMDB

Prerequisites

The SMW HA system can be configured to store the Power Management Database (PMDb) on shared storage, a logical disk configured as a LUN (Logical Unit) or logical volume on the boot RAID.

IMPORTANT: Cray strongly recommends using mirrored storage, if available, for the PMDB; for more information, see [Storage for the Power Management Database \(PMDb\)](#) on page 9. To move the PMDB from shared storage to mirrored storage, see [Migrate PMDB Data from the Boot RAID to Mirrored Storage](#) on page 322.

Before beginning this procedure:

- Ensure that the boot RAID contains a LUN for the PMDB with sufficient space for the data. Use the following command to check the size of `/var/lib/pgsql` on the local disk:

```
smw1:~ # du -hs /var/lib/pgsql
```

- Check that the boot RAID is connected.
- Ensure that the SMW HA software is correctly configured and that the HA cluster is running correctly.
- To capture typescript output from this procedure, do not use a typescript session running directly on the SMW. To save the output of this procedure, use the `script` command to start the typescript session on your local workstation before logging into the SMW, as in this example:

```
workstation> script -af my_output_file
Script started, file is my_output_file
workstation> ssh crayadm@smw1
```

Use this procedure to configure the RAID disk and transfer the power management data base (PMDB) to the power management disk on the shared boot RAID.

1. Shut down the Cray system by typing the following command as `crayadm` on the active SMW (`smw1`).

```
crayadm@smw1:~>xtbootsys -s last -a auto.xtshutdown
```

2. Log into the active SMW as `root`, either at the console or by using the actual (not virtual) host name.

IMPORTANT: You must log in directly as `root`. Do not use `su` from a different SMW account such as `crayadm`.

3. Change to the directory containing the `SMWHAconfig` command.

```
smw1:~ # cd /opt/cray/ha-smw/default/hainst
```

4. Use the `SMWHAconfig` command to move the PMDB and configure the required HA resources. In the following command, replace `scsi-xxxxxxxx` with the persistent device name for the PMDB directory on the boot RAID.

```
smw1:~ # ./SMWHAconfig --add_disk=pm-fs \
--device=/dev/disk/by-id/scsi-xxxxxxxx --directory=/var/lib/pgsql
```

This command mounts the PMDB directory (`/var/lib/pgsql`) to the boot RAID, copies the PMDB data, and configures the HA resources `pm-fs` and `postgresqld`.

5. Reboot `smw1` and wait for the reboot to finish.

```
smw1:~ # reboot
```

Before continuing, wait until `smw1` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw1`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw1` is online.

6. Reboot `smw2` and wait for the reboot to finish.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

7. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):      Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons   (lsb:rsms):                    Started smw1
stonith-1     (stonith:external/ipmi):      Started smw2
stonith-2     (stonith:external/ipmi):      Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):   Started smw1
  cray-syslog  (lsb:cray-syslog):            Started smw1
  homedir      (ocf::heartbeat:Filesystem):   Started smw1
  md-fs        (ocf::heartbeat:Filesystem):   Started smw1
  pm-fs        (ocf::heartbeat:Filesystem):   Started smw1
  postgresql   (lsb:postgresql):            Started smw1
  mysqld       (ocf::heartbeat:mysql):       Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-`root` user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
```

```

Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
...
Cleaning resource on node=smw2 for resource=Notification

```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

8. Verify that the Power Management Database is on the boot RAID and that the required PMDB resources are running.
 - a. Examine the log file `/opt/cray/ha-smw/default/hainst/SMWHAconfig.out` to verify that the Power Management Database disk appears in the Cluster RAID Disks section (at the end of the file), as in this example.

```

----- Cluster RAID Disks -----
07-07 20:47 INFO  MySQL Database disk = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  Log disk             = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  /home disk          = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  PM database disk    = /dev/disk/by-id/scsi-360080e5..xxx
07-07 20:47 INFO  ***** Ending of HA software add_disk *****

```

- b. Ensure that the power management file system is mounted by checking for `/var/lib/pgsql` in the output of the `df` command.

```

smw1:~ # df

```

Filesystem	1K-blocks	Used	Available	Use%	Mounted on
/dev/sda2	120811676	82225412	32449332	72%	/
udev	16433608	756	16432852	1%	/dev
tmpfs	16433608	37560	16396048	1%	/dev/shm
/dev/sdo	483807768	197536596	261695172	44%	/var/opt/cray/disk/1
/dev/sdp	100791728	66682228	28989500	70%	/home
/dev/sdq	100791728	484632	95187096	1%	/var/lib/mysql
/dev/sdr	30237648	692540	28009108	3%	/var/lib/pgsql

- c. Check the output of `crm_mon` to ensure that the `pm-fs` and `postgresqd` resources are running.

```

smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.

```

```

Online: [ smw1 smw2 ]

```

```

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):       Started smw1
dhcpd (lsb:dhcpd):      Started smw1

```

```

fsync (ocf::smw:fsync):      Started smw1
hss-daemons (lsb:rsms):      Started smw1
stonith-1 (stonith:external/ipmi):      Started smw2
stonith-2 (stonith:external/ipmi):      Started smw1
Resource Group: HSSGroup
  ml-fs (ocf::heartbeat:Filesystem):      Started smw1
  cray-syslog (lsb:cray-syslog):      Started smw1
  homedir (ocf::heartbeat:Filesystem):      Started smw1
  md-fs (ocf::heartbeat:Filesystem):      Started smw1
  pm-fs (ocf::heartbeat:Filesystem):      Started smw1
  postgresql (lsb:postgresql):      Started smw1
  mysqld (ocf::heartbeat:mysql): Started smw1

```

Migrate PMDB Data from the Boot RAID to Mirrored Storage

Prerequisites

Before beginning this procedure:

- Ensure that the mirrored PMDB disk has been configured as specified in [Configure Mirrored Storage with DRBD for the PMDB](#) on page 172.
- Identify the device name of the boot RAID partition containin the Power Management Database (PDMB).

Use the following procedure to move the PMDB data from shared storage on the boot RAID to the mirrored storage on the DRBD disk.

- Log into the active SMW as `root`.
- Put the cluster in maintenance mode.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- Stop `rsms`.

```
smw1:~# rsms stop
```

- Stop `postgresql`.

```
smw1:~# /etc/init.d/postgresql stop
```

- Mount the boot RAID partition previously used by the PMDB.

```
smw1:~# mount boot_RAID_partition /mnt/pgsql_tmp
```

- Back up the existing copy of `/var/lib/pgsql`, if possible.

```
smw1:~# cp -pr /var/lib/pgsql /var/lib/pgsql-backup
```

- Remove the existing contents of `/var/lib/pgsql` on the mirrored disk.

```
smw1:~# rm -rf /var/lib/pgsql/*
```

- Copy the PMDB contents from the boot RAID partition to `/var/lib/pgsql`.

```
smw1:~# cp -pr /mnt/pgsql_tmp/* /var/lib/pgsql
```

9. Start postgresql.

```
smw1:~# /etc/init.d/postgresql start
```

10. Check the postgresql status.

```
smw1:~# /etc/init.d/postgresql status
Checking for PostgreSQL
9.1.12: running
```

11. Start rsms.

```
smw1:~# rsms start
```

12. Inspect the status of the rsms daemons and the contents of /var/opt/cray/log/power_management-YYYYMMDD, where YYYYMMDD is today's date. If xtpmd is running and no database errors are noted, the transfer went properly.

```
smw1:~# rsms status
cluster is in maintenance mode and daemons are not under cluster control
Checking for RSMS service:
erd.. running
Checking for RSMS service:
erdh.. running
Checking for RSMS service:
sm.. running
Checking for RSMS service:
nm.. running
Checking for RSMS service:
bm.. running
Checking for RSMS service:
sedc_manager.. running
Checking for RSMS service:
cm.. running
Checking for RSMS service:
xtpmd.. running
Checking for RSMS service:
erfsd.. running
Checking for RSMS service:
xtremoted.. running
```

13. If the rsms status is good, remove the backup of /var/lib/pgsql.

14. Wait for the PMDB to sync completely. A full initial synchronization takes a long time, regardless of the size of the PMDB. The time to synchronize a 1 TB external DRBD disk is approximately 10 hours. Check the status of the DRBD sync operation with either rcdbrd status or cat /proc/drbd. The rcdbrd output is easier to read, but /proc/drbd contains more status information and includes an estimate of time to completion.

```
smw1:~# rcdbrd status
drbd driver loaded OK; device status:
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
m:res cs          ro          ds          p mounted
```

```
fstype
0:r0 SyncSource Primary/Secondary UpToDate/Inconsistent C /var/lib/pgsql
ext3
... sync'ed: 72.7% (252512/922140)M
```

```
smw1:~# cat /proc/drbd
version: 8.4.4 (api:1/proto:86-101)
GIT-hash: 599f286440bd633d15d5ff985204aff4bccffadd build by phil@fat-tyre,
2013-10-11 16:42:48
 0: cs:SyncSource ro:Primary/Secondary ds:UpToDate/Inconsistent C r-----
    ns:695805444 nr:12508 dw:1808112 dr:694131606 al:171 bm:43068 lo:0 pe:2 ua:
0 ap:0 ep:1 wo:f oos:260636656
    [=====>.....] sync'ed: 72.4% (254524/922140)M
    finish: 2:21:07 speed: 30,768 (29,720) K/sec
```

15. Take the cluster out of maintenance mode.

```
smw1:~# crm configure property maintenance-mode=false 2 > /dev/null
```

Troubleshooting SMW HA Installation Problems

Use the following procedures for serious installation problems.

- [Restore a Previous SMW HA Configuration](#) on page 399: If there are problems with an SMW HA update, use this procedure to restore the previous configuration.
- [Disable the SMW HA Configuration](#) on page 400: How to convert an SMW HA cluster to a system with a single stand-alone SMW.
- [Re-enable the SMW HA Configuration](#) on page 403: How to restore the SMW HA configuration to a disabled cluster.
- [Migrate PMDB Data from Mirrored Storage to the Boot RAID](#) on page 406: If there are problems with mirrored storage via DRBD, use this procedure to move the PMDB to the shared boot RAID.

Restore a Previous SMW HA Configuration

The `SMWHAconfig` command backs up the cluster configuration before beginning an update. If there are serious problems with the update, use this procedure to restore the previous configuration.



CAUTION: If you reinstall the SMW HA software on an existing SMW HA cluster, `SMWHAconfig` automatically deletes any existing data in the shared directories on the boot RAID. When redoing an initial installation, existing data is not reused.

1. Log in as `root` on the active SMW (`smw1`).
2. Put both SMWs in standby mode.

NOTE: Replace `smw1` with the host name of the active SMW. Replace `smw2` with the host name of the passive SMW.

```
smw1:~ # crm node standby smw1
smw1:~ # crm node standby smw2
```

3. Erase the cluster resources.

```
smw1:~ # crm configure erase
```

4. Locate the previous configuration in the directory `/opt/cray/ha-smw/default/hainst`, in a file named `_CLUSTER_CONFIG_BACKUP_YYYY-MM-DD-hh:mm` (for example, `_CLUSTER_CONFIG_BACKUP_2013-06-11-10:11`).
5. Restore the previous configuration. In the following command, replace `YYYY-MM-DD-hh:mm` with the actual timestamp of the file.

```
smw1:~ # crm configure load replace \
/opt/cray/ha-smw/default/hainst/_CLUSTER_CONFIG_BACKUP_YYYY-MM-DD-hh:mm 2> /dev/null
.
.
.
```

6. Take both SMWs out of standby mode (put them back online).

```
smw1:~ # crm node online
```

```
smw2:~ # crm node online
```

Disable the SMW HA Configuration

If problems occur during system configuration and testing, it may be helpful to temporarily disable the SMW HA cluster without uninstalling HA cluster software, then re-enable the cluster after fixing the problems. You can disable an SMW HA cluster by converting it to two unclustered SMWs. The active SMW is converted to a stand-alone SMW. The passive SMW is powered off to prevent interference between the two SMWs.

The following information is required for this procedure:

- DRAC IP address of the passive SMW (see [Site-dependent Configuration Values for an SMW HA System](#)).
- Virtual host name and virtual IP address of the cluster (see [Site-dependent Configuration Values for an SMW HA System](#)). The examples in this procedure use the virtual host name `virtual-smw`; substitute the actual host name for the system.
- Host names of the active and passive SMWs (see [Site-dependent Configuration Values for an SMW HA System](#)). The examples in this procedure use the host names `smw1` and `smw2`; substitute the actual host names for the system.
- IP addresses of the original (virtual) Ethernet ports (see [Fixed IP Addresses for an SMW HA system](#) on page 16).
- Device names of the local MySQL database, Log directory, and home directories. You will need the by-path device names for the following directories on the local disk:
 - `/var/lib/mysql`
 - `/var/opt/cray/disk/1`
 - `/home`
 - `/var/lib/pgsql` (if the PMDB is on the shared boot RAID)
- Persistent device names of the shared directories on the boot RAID.
 - `/var/lib/mysql` (MySQL database)
 - `/var/opt/cray/disk/1` (Log directory)
 - `/home` (home directories)
 - `/var/lib/pgsql` (if the PMDB is on the shared boot RAID)

For more information, see [Shared Storage on the Boot RAID](#) on page 8.

TIP: Execute this command as root to display the configured device names.

```
smw1:~ # crm configure show | grep device | awk '{print $2 " " $3}' | sed 's/"//g'
device=/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9c01 directory=/home
device=/dev/disk/by-id/scsi-360080e500023bff6000006b3515d9bdf directory=/var/lib/mysql
device=/dev/disk/by-id/scsi-360080e500023bff6000006b1515d9bc9 directory=/var/opt/cray/disk/1
device=/dev/disk/by-id/scsi-360080e500023bff6000006b5515d9d01 directory=/var/lib/pgsql
```

The examples in this procedure show the host names `virtual-smw` (virtual host name for the cluster), `smw1` (active SMW), and `smw2` (passive SMW). Substitute the actual host names for the system.

1. Log on to the active SMW (`smw1`) as `root`.
2. Put both SMWs in standby mode.

NOTE: Replace `smw1` with the host name of the active SMW. Replace `smw2` with the host name of the passive SMW.

```
smw1:~ # crm node standby smw1
smw1:~ # crm node standby smw2
```

3. On `smw1`, edit the `/etc/sysconfig/network/ifcfg-eth*` files to specify the original IP addresses for the Ethernet ports (`eth0`, `eth1`, `eth2`, `eth3`, and `eth4`).

NOTE: This step assumes that the site uses the standard fixed IP addresses for these ports. If the site uses different IP addresses, ensure that the final digit in the dotted quad is 1 (the virtual IP address), not 2 or 3.

- a. In `/etc/sysconfig/network/ifcfg-eth0`, change the `IPADDR` value to the virtual IP address of the SMW (for example, the IP address for `virtual-smw.cray.com`).
- b. In `/etc/sysconfig/network/ifcfg-eth1`, change the `IPADDR` value to `'10.1.1.1/16'`.
- c. In `/etc/sysconfig/network/ifcfg-eth2`, change the `IPADDR` value to `'10.2.1.1/16'`.
- d. In `/etc/sysconfig/network/ifcfg-eth3`, change the `IPADDR` value to `'10.3.1.1/16'`.
- e. In `/etc/sysconfig/network/ifcfg-eth4`, change the `IPADDR` value to `'10.4.1.1/16'`.

4. Disable the SMW HA service configuration.

IMPORTANT: The command order is important. Do not change the order of these commands.

```
smw1:~ # chkconfig openais off
smw1:~ # chkconfig mysql on
smw1:~ # chkconfig xinetd off
smw1:~ # chkconfig dbMonitor on
smw1:~ # chkconfig rsms on
smw1:~ # chkconfig dhcpcd on
smw1:~ # chkconfig postgresql on
```

5. Power off the passive SMW (`smw2`). For `drac-ip-address`, specify the passive SMW's iDRAC IP address.

```
smw2:~ # /usr/bin/ipmitool -I lanplus -U root -H drac-ip-address -a chassis power off
```

NOTE: Enter the `root` password when prompted.

6. Edit `/etc/hosts` to replace the active SMW's host name and IP address with the virtual values, so that users can use the same name to access the system. (For example, change `smw1` to `virtual-smw`.)
 - a. Locate the line that specifies the active SMW's IP address and host name, as in this example:

```
172.30.49.161 smw1 virtual-smw1
```

- b. Change this line to the virtual IP address and host name of the cluster, as in this example:

```
172.30.49.160 smw virtual-smw
```

7. Execute the following commands to update `/etc/HOSTNAME` with the virtual host name for the cluster.

NOTE: This example shows the host names `smw1.us.cray.com` and `virtual-smw.us.cray.com`. Substitute the full host name for the cluster.

```
virtual-smw:~# cat /etc/HOSTNAME
smw1.us.cray.com
virtual-smw:~# echo virtual-smw.us.cray.com > /etc/HOSTNAME
```

8. For each shared directory, check whether the shared RAID disk is mountable. In the following commands, replace `scsi-xxxxxx` with the persistent device name for a shared directory on the boot RAID. .

```
smw1:~ # mkdir -p /mnt/test
smw1:~ # mount /dev/disk/by-id/scsi-deviceA /mnt/test
smw1:~ # echo $?
0
smw1:~ # umount /mnt/test
```

If the `echo` command displays the value 1 (as the error status for the `mount` command), the shared RAID disk is not mountable. If this is the case, define the mount points in `/etc/fstab` for the MySQL database and the Log directory on local disk.

- a. Edit `/etc/fstab` and locate the lines containing the by-path device names for `/var/lib/mysql` and `/var/opt/cray/disk/1`. These lines are commented out in a cluster system.

For example, locate the following lines:

```
# /dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0-part1 /var/opt/cray/disk/1 ...
# /dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0-part1 /var/lib/mysql ...
```

- b. Remove the comment character from these lines, as in this example:

```
/dev/disk/by-path/pci-0000:05:00.0-sas-phy4-0x4433221104000000-lun-0-part1 /var/opt/cray/disk/1 ...
/dev/disk/by-path/pci-0000:05:00.0-sas-phy5-0x4433221105000000-lun-0-part1 /var/lib/mysql ...
```

9. If the shared RAID disk is mountable, define the mount points in `/etc/fstab` for the MySQL database and the Log directory on the boot RAID.

- a. Edit `/etc/fstab` to add the RAID disk names.
b. Change the permissions of directory `/var/lib/mysql/hssds` to `mysql`.

```
smw1:~ # chgrp -R mysql /var/lib/mysql/hssds
smw1:~ # chown -R mysql /var/lib/mysql/hssds
```

10. Reboot the SMW.

```
smw1:~ # reboot
```

The formerly active SMW now functions as a standard, unclustered SMW with the cluster's virtual host name (for example, `virtual-smw`). The other (formerly passive) SMW must remain powered off because it is still configured for the SMW HA cluster.

Re-enable the SMW HA Configuration

To re-enable a disabled SMW HA cluster, undo the changes to the active SMW (`smw1`) that were made in [Disable the SMW HA Configuration](#). The two SMWs will be returned to the active/passive configuration for the SMW HA cluster.

The following information is required for this procedure:

- Virtual host name and virtual IP address of the cluster (see [Site-dependent Configuration Values for an SMW HA System](#)).
- Host names of the active and passive SMWs (see [Site-dependent Configuration Values for an SMW HA System](#)).
- IP addresses of the cluster-specific Ethernet ports for the active and passive SMW (see [Fixed IP Addresses for an SMW HA system](#) on page 16).

The examples in this procedure show the host names `virtual-smw` (virtual host name for the cluster), `smw1` (active SMW), and `smw2` (passive SMW). Substitute the actual host names for the system.

1. Log in as `root` to the running SMW (for example, `virtual-smw`).
2. Edit the `/etc/sysconfig/network/ifcfg-eth*` files to restore the cluster-specific IP addresses for the Ethernet ports (`eth0`, `eth1`, `eth2`, `eth3`, and `eth4`).

NOTE: This step assumes that the site uses the standard fixed IP addresses for these ports. If the site uses different IP addresses, ensure that the final digit in the dotted quad is 2 or 3, not 1 (the virtual SMW).

- a. Identify the final digit in the dotted quad of IP addresses for HA cluster.

```
virtual-smw:~# egrep -e '(smw.*smw-net1|smw-net1.*smw)' /etc/hosts /etc/hosts | \
awk '{print $1}' | awk -F"." '{print $4}'
2
```

NOTE: The returned value is usually 2, which means that the active SMW is `smw1`. The value 3 means that `smw2` is the active SMW; if so, use 3 instead of 2 as the final digit of the dotted quad in the following substeps.

- b. In `/etc/sysconfig/network/ifcfg-eth1`, change the `IPADDR` value to `'10.1.1.2/16'`.
 - c. In `/etc/sysconfig/network/ifcfg-eth2`, change the `IPADDR` value to `'10.2.1.2/16'`.
 - d. In `/etc/sysconfig/network/ifcfg-eth3`, change the `IPADDR` value to `'10.3.1.2/16'`.
 - e. In `/etc/sysconfig/network/ifcfg-eth4`, change the `IPADDR` value to `'10.4.1.2/16'`.
 - f. In `/etc/sysconfig/network/ifcfg-eth0`, change the `IPADDR` value to the cluster's virtual IP address. For example, if the actual IP address for `smw1` is `172.30.49.61`, change this value to `172.30.49.60`.
3. Restore the SMW HA service configuration.

IMPORTANT: The command order is important. Do not change the order of these commands.

```
virtual-smw:~# chkconfig rsms off
virtual-smw:~# chkconfig dbMonitor off
```

```
virtual-smw:~# chkconfig xinetd on
virtual-smw:~# chkconfig openais on
virtual-smw:~# chkconfig mysql off
virtual-smw:~# chkconfig dhcpd off
```

4. Edit `/etc/hosts` to change the SMW's host name and IP address.

- a. Locate the line that specifies the cluster's virtual IP address and host name, as in this example:

```
172.30.49.160    smw virtual-smw
```

- b. Change this line to the actual IP address and host name of `smw1`, as in this example:

```
172.30.49.161    smw1 virtual-smw1
```

5. Execute the following commands to update `/etc/HOSTNAME` with the actual name for `smw1`.

NOTE: This example shows the host names `virtual-smw.us.cray.com` and `smw1.us.cray.com`. Substitute the full host name for `smw1`.

```
virtual-smw:~# cat /etc/HOSTNAME
virtual-smw.us.cray.com
virtual-smw:~# echo smw1.us.cray.com > /etc/HOSTNAME
```

6. Edit `/etc/fstab` to remove the mount points for directories `/var/lib/mysql` (MySQL directory), `/var/opt/cray/disk/1` (Log directory), and `/home` (home directories).

- a. If the shared RAID disks are used, remove the mount points for these devices. For more information, see [Disable the SMW HA Configuration](#) on page 400.
- b. If the local disks are used, comment out (add a comment character to) each line that defines a local disk mount point. For more information, see [Disable the SMW HA Configuration](#) on page 400.

7. Change the permission of the `/var/lib/mysql/hssds` directory to `root`.

```
virtual-smw:~# chgrp -R root /var/lib/mysql/hssds
virtual-smw:~# chown -R root /var/lib/mysql/hssds
```

8. Reboot the SMW and wait for it to finish rebooting.

```
virtual-smw:~# reboot
```

Before continuing, wait until the SMW has rejoined the cluster. After the SMW responds to a `ping` command, log into the SMW, sleep for at least 2 minutes, then execute the `crm_mon -1` command to verify that the SMW is online.

This SMW is now the active SMW in the cluster (`smw1`).

9. Power on the second SMW (`smw2`) and wait for it to finish rebooting.

```
smw2:~ # reboot
```

Before continuing, wait until `smw2` has rejoined the cluster. After the SMW responds to a `ping` command, log into `smw2`, sleep for at least 2 minutes, then execute the `crm_mon -r1` command to verify that `smw2` is online.

10. Verify that all resources are running.

- a. Display the cluster status.

```
smw1:~ # crm_mon -r1
Last updated: Mon Oct 27 01:19:23 2014
Last change: Thu Oct 23 15:15:04 2014 by root via crm_attribute on smw2
Stack: classic openais (with plugin)
Current DC: smw1 - partition with quorum
Version: 1.1.9-2db99f1
2 Nodes configured, 2 expected votes
19 Resources configured.
```

```
Online: [ smw1 smw2 ]
```

```
ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP3     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP4     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterMonitor (ocf::smw:ClusterMonitor):     Started smw1
Notification   (ocf::heartbeat:MailTo):      Started smw1
dhcpd          (lsb:dhcpd):                   Started smw1
fsync          (ocf::smw:fsync):              Started smw1
hss-daemons    (lsb:rsms):                    Started smw1
stonith-1      (stonith:external/ipmi):      Started smw2
stonith-2      (stonith:external/ipmi):      Started smw1
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):   Started smw1
  cray-syslog  (lsb:cray-syslog):            Started smw1
  homedir     (ocf::heartbeat:Filesystem):   Started smw1
  md-fs       (ocf::heartbeat:Filesystem):   Started smw1
  pm-fs       (ocf::heartbeat:Filesystem):   Started smw1
  postgresql  (lsb:postgresql):            Started smw1
  mysqld      (ocf::heartbeat:mysql):       Started smw1
```

Note that `crm_mon` may display different resource names, group names, or resource order on the system.

- b. Examine the `crm_mon` output. Verify that each resource has started by looking for `Started smw1` or `Started smw2`. Also look for any failed actions at the end of the output.
- c. If not all resources have started or if any failed actions are displayed, execute the `clean_resources` command on either SMW.

IMPORTANT: When running the `clean_resources` command, you must be directly logged in as `root` (instead of using `su` from a `crayadm` login), because `clean_resources` terminates all non-root user sessions.

```
smw1:~ # clean_resources
Cleaning resources on node smw1
Cleaning resource on node=smw1 for resource=stonith-1
Cleaning resource on node=smw1 for resource=stonith-2
Cleaning resource on node=smw1 for resource=dhcpd
Cleaning resource on node=smw1 for resource=cray-syslog
Cleaning resource on node=smw1 for resource=ClusterIP
Cleaning resource on node=smw1 for resource=ClusterIP1
Cleaning resource on node=smw1 for resource=ClusterIP2
...
Cleaning resources on node smw2
Cleaning resource on node=smw2 for resource=stonith-1
Cleaning resource on node=smw2 for resource=stonith-2
```

```
...
Cleaning resource on node=smw2 for resource=Notification
```

After running `clean_resources`, wait several minutes for cluster activity to settle. You can check cluster status with the `crm_mon -r1` command. If the output of this command shows only a subset of the SMW HA services, wait for another minute, then check again. For more information, see the `clean_resources(8)` man page.

11. If necessary, restore the original `postgresql` service state on `smw1`.
 - a. Check the original PMDB configuration on the passive SMW (`smw2`).

```
2~ # chkconfig postgresql
```

- b. If the service state on `smw1` is different, change the state on this system to match the state on `smw2`.

The SMW HA configuration is now restored. To verify that the SMW HA cluster is running correctly, see [Verify the SMW HA Cluster Configuration](#) on page 181.

Migrate PMDB Data from Mirrored Storage to the Boot RAID

Prerequisites

NOTE: Mirrored storage with the Distributed Replicated Block Device (DRBD) is preferred for the Power Management Database (PMDB). Do not move the PMDB to the boot RAID unless no other option is available.

Before beginning this procedure:

- Do not disable DRBD in the SMW HA configuration or physically remove the mirrored disks from the SMWs. The mirrored DRBD storage must remain accessible for this procedure.
- Identify temporary storage with sufficient space for `/var/lib/pgsql`, such as the root partition on the SMW or an external storage device.

Use the following procedure to move the PMDB data from mirrored storage to shared storage on the boot RAID.

1. As `crayadm`, shut down the Cray system.

```
crayadm@smw1:~> xtbootsys -s last -a auto.xtshutdown
```

2. Log into the active SMW as `root`.
3. Put the cluster in maintenance mode.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

4. Stop `rsms`.

```
smw1:~# rsms stop
```

5. Stop `postgresql`.

```
smw1:~ # /etc/init.d/postgresql stop
```

- Copy the contents of `/var/lib/pgsql` to temporary storage, preserving permissions and ownership. The following example shows how to copy `/var/lib/pgsql` to a directory on the root partition on the SMW (for example, `/pgsql_tmp`).

```
smw1:~# mkdir /pgsql_tmp
smw1:~# cp -pr /var/lib/pgsql/* /pgsql_tmp
smw1:~# ls -l /pgsql_tmp/
total 12
drwx----- 14 postgres postgres 4096 Mar 13 12:14 data
-rw-r--r-- 1 postgres postgres 1224 Feb  4 17:57 initlog
drwx----- 2 root      root      4096 Feb  6 14:35 lost+found
```

- Remove the DRBD disk from the SMW HA configuration as described in [Remove the Mirrored Storage Disk for the PMDB](#) on page 408.
- Add the RAID disk to the SMW HA configuration as described in [Configure Shared Storage on the Boot RAID for the PMDB](#) on page 177.
- Put the cluster in maintenance mode.

```
smw1:~# crm configure property maintenance-mode=true 2> /dev/null
```

- Stop `rsms`.

```
smw1:~# rsms stop
```

- Stop `postgresql`.

```
smw1:~# /etc/init.d/postgresql stop
```

- Remove the existing contents of `/var/lib/pgsql` on the boot RAID.

```
smw1:~# rm -rf /var/lib/pgsql/*
```

- Copy the contents of `/var/lib/pgsql` from the temporary location (see step 6 on page 407) to the boot RAID partition. The following example assumes that `/pgsql_tmp` was used as temporary storage.

```
smw1:~# cp -rp /pgsql_tmp/* /var/lib/pgsql
```

- Start `postgresql`.

```
smw1:~# /etc/init.d/postgresql start
```

- Check the `postgresql` status.

```
smw1:~# /etc/init.d/postgresql status
Checking for PostgreSQL
9.1.12:                                     running
```

- Start `rsms`.

```
smw1:~# rsms start
```

17. Inspect the status of the `rsms` daemons and the contents of `/var/opt/cray/log/power_management-YYYYMMDD`, where `YYYYMMDD` is today's date. If `xtpmd` is running and no database errors are noted, the transfer went properly.

```
smw1:~# rsms status
cluster is in maintenance mode and daemons are not under cluster control
Checking for RSMS service:
erd..                                running
Checking for RSMS service:
erdh..                               running
Checking for RSMS service:
sm..                                  running
Checking for RSMS service:
nm..                                  running
Checking for RSMS service:
bm..                                  running
Checking for RSMS service:
sedc_manager..                       running
Checking for RSMS service:
cm..                                  running
Checking for RSMS service:
xtpmd..                               running
Checking for RSMS service:
erfsd..                              running
Checking for RSMS service:
xtremoted..                          running
```

18. Take the cluster out of maintenance mode.

```
smw1:~# crm configure property maintenance-mode=false 2 > /dev/null
```

Remove the Mirrored Storage Disk for the PMDB

Cray recommends using mirrored storage with the Distributed Replicated Block Device (DRBD) for the power management database (PMDb). However, if mirrored storage must be disabled, use this procedure to remove the mirrored storage disk from the SMW HA configuration.

1. Log in as root on the first SMW (`smw1`).

```
smw1:~# ssh root@smw1
```

2. Navigate to the directory containing the `SMWHAconfig` command.

```
smw1:~# cd /opt/cray/ha-smw/default/hainst
```

3. Execute `SMWHAconfig` to remove the DRBD disk.

```
smw1:~# ./SMWHAconfig --remove_disk=pm-fs
```

4. Stop the DRBD service on both `smw1` and `smw2`.

```
smw1:~# rcdbrd stop
```

```
smw2:~# rcdbrd stop
```

5. Reboot the active SMW (*smw1*) and wait for it to boot completely.
6. Reboot the other SMW (*smw2*) and wait for it to boot completely.
7. Examine the output of `crm status` to ensure that there are no entries for `postgresd`, `ip_drbd_pgsql`, and the `Masters` and `Slaves` resources.

```
smw1:~# crm status
Last updated: Mon Jan 26 14:20:16 2015
Last change: Thu Jan 15 10:44:11 2015
Stack: corosync
Current DC: smw2 (167903490) - partition with quorum
Version: 1.1.12-ad083a8
2 Nodes configured
18 Resources configured

Online: [ smw2 smw1 ]

ClusterIP      (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP1     (ocf::heartbeat:IPaddr2):      Started smw1
ClusterIP2     (ocf::heartbeat:IPaddr2):      Started smw1
.
.
Resource Group: HSSGroup
  ml-fs        (ocf::heartbeat:Filesystem):    Started smw1
  cray-syslog  (systemd:llmrtd.service):      Started smw1
  homedir     (ocf::heartbeat:Filesystem):    Started smw1
  md-fs       (ocf::heartbeat:Filesystem):    Started smw1
  mysqld      (ocf::heartbeat:mysql):        Started smw1
```