XC™ Series CLE 5.2 to CLE 6.0 Software Migration Overview (CLE 6.0.UP03) S-2574

Contents

1 About XC™ Series CLE 5.2 to CLE 6.0 Software Migration Overview

Scope and Audience

The *XC™ Series CLE 5.2 to CLE 6.0 Software Migration Overview* (S-2574) provides an overview of the tasks necessary to migrate from CLE 5.2.UP04 / SMW 7.2.UP04 software to CLE 6.0.UP03 / SMW 8.0.UP03 software on Cray XC™ series hardware. Although the XE/XK series hardware could run CLE 5.2.UP04 / SMW 7.2.UP04 software, that hardware is not supported with CLE 6.0 / SMW 8.0 releases.

This publication does not include detailed migration procedures; for those, see the following:

- *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Virtual SMW* (S-2575)
- *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Physical SMW* (S-2580)
- *XC™ Series esLogin to eLogin Migration Guide* (S-2584)

This publication is intended for sites that wish to migrate from CLE 5.2.UP04 / SMW 7.2.UP04 software to CLE 6.0.UP03 / SMW 8.0.UP03 software on Cray XC™ series hardware.

CLE 6.0.UP03 / SMW 8.0.UP03 Release

XC™ Series CLE 5.2 to CLE 6.0 Software Migration Overview (CLE 6.0.UP03) S-2574 supports Cray software release CLE 6.0.UP03 / SMW 8.0.UP03 for Cray XC™ Series systems, released on 16 February 2017. This is the initial release of this publication.

Feedback

Your feedback is important to us. 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 comments to pubs@cray.com.

Trademarks

2 Introduction to Software Migration from CLE 5.2 to CLE 6.0

The Cray CLE 6.0 / SMW 8.0 releases use SUSE® Linux Enterprise Server (SLES) 12 as the base operating system on the SMW. The most recent prior release, CLE 5.2.UP04 / SMW 7.2.UP04, was based on SLES 11. Because SLES 12 represents a major change in architecture and features, the transition from SLES 11 to SLES 12 requires a software *migration* rather than a software *upgrade*.

To help with this, Cray has created a one-time CLE/SMW migration process and esLogin migration process for use by customer staff and on-site Cray field support staff to migrate from CLE 5.2.UP04 / SMW 7.2.UP04 (SLES 11) to CLE 6.0.UP03 / SMW 8.0.UP03 (SLES 12).

Migration Goal and the "Migration SMW"

The goal of the migration process is to minimize system downtime (unavailability to run user jobs) while preserving necessary configuration and operation data. To achieve this goal, the CLE/SMW process requires the use of a *migration SMW*, which can be either a virtual SMW hosted on an on-site Cray laptop or a spare physical SMW and boot RAID that are not connected to the currently running XC system.

Migration Documentation

This publication provides an overview of the CLE/SMW migration process. Details are available in

- *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Virtual SMW* (S-2575)
- *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Physical SMW* (S-2580)

For information about migrating external login nodes, see *XC™ Series esLogin to eLogin Migration Guide* (S-2584).

Migration Scope

- **Software Releases.** Migration of CLE and SMW software is supported from CLE 5.2.UP04 and SMW 7.2.UP04 to CLE 6.0.UP03 and SMW 8.0.UP03 only. To migrate from an earlier release, update to CLE 5.2.UP04 and SMW 7.2.UP04 first, then use this migration process. To migrate to a later CLE 6.0 / SMW 8.0 release, use this migration process, then update from CLE 6.0.UP03 and SMW 8.0.UP03.
- **Hardware.** This migration process applies to XC series systems only. CLE 6.0 / SMW 8.0 releases do not support XE or XK series hardware. Also, the CLE 6.0.UP03 / SMW 8.0.UP03 release does not support Intel® Xeon Phi™ "Knight's Corner" (KNC) nodes. To reduce risk, this software migration process assumes that no hardware upgrades will be attempted during the migration.
- **CIMS/esLogin.** Migration is supported from a CIMS (Cray Integrated Management Services) running Bright Computing software to manage esLogin-based CDL (Cray Development and Login) nodes to a CMC (Cray Management Controller) running CSMS (Cray System Management Software) and OpenStack software to manage eLogin-based CDL nodes. The process to extract information from a CDL node managed by a CIMS running Bright Computing software is provided in *XC™ Series esLogin to eLogin Migration Guide* (S-2584).

- **SMW HA.** Migration of an SMW HA system is supported from SLEHA11.SP3.UP02 to SLEHA12.SP0.UP03 in two stages: migrate the first SMW using this migration process, and then proceed with the standard SMW HA fresh install process for installing and configuring SMW HA software on the first SMW, setting up the second SMW, and configuring the SMW HA cluster, as described in *XC™ Series SMW HA Installation Guide* (S-0044). The detailed migration guides contain some "SMW HA only" steps and notes that apply to migration of the first SMW.
- **Power Management.** This migration process does not migrate power management data. See *XC™ Series Power Management Administration Guide (CLE 6.0.UP03) S-0043* for information pertaining to backing up the database. The erfs data must be regenerated.

Migration Process Phases

1. **Training.** This first phase consists of learning about the new Cray Management System (CMS) through Cray training classes, Cray technical publications, and publicly available Ansible documentation (Ansible is leveraged heavily by the new management system_.
2. **Planning.** This second phase plans the hardware and software configurations needed for a successful migration. Sites are encouraged to contact their district service manager for help with this phase.
3. **Preparation of Physical Migration SMW and Boot RAID.** This phase applies only if using a physical migration SMW as opposed to a virtual one. It performs a fresh install of the CLE 6.0.UP03 / SMW 8.0.UP03 software release on the spare physical SMW that will be used for migration.
4. **Preparation of Configuration Data and Images.** This phase extracts configuration data from the currently running system and transfers that data to configuration worksheets in the new release. It also builds the appropriate software images and assigns them to nodes. Most of the procedures in this phase are done on the migration SMW.
5. **Preservation of Other Data.** This phase captures accounting, operational, and user data from the currently running CLE system just prior to shutting it down.
6. **Shutdown and Switch.** This final phase has two different scenarios.

   virtual If a virtual migration SMW is used, this phase performs a fresh install of all software on the original SMW, which is connected to the XC hardware. It then moves the configuration data and images from the virtual migration SMW to the original SMW, which now has CLE 6.0 / SMW 8.0 software installed.

   physical If a physical migration SMW is used, this phase disconnects the original SMW and boot RAID from the XC hardware and connects the migration SMW and new boot RAID to the XC hardware.

   In both cases, this phase also completes any configuration requiring connection to XC hardware.

   NOTE: Rolling back to the CLE 5.2 / SMW 7.2 system at this point in the migration process is possible only if using a physical migration SMW.

Migration Caveats

The following list of features and components have one or more associated caveats to be aware of.

SMW HA When doing a migration where the end result is SMW HA, the two SMWs that will run SMW 8.0 / CLE 6.0 with SLEHA12.SP0.UP03 must be matched hardware:

same model (both Dell PowerEdge™ R815 Rack Servers or both Dell PowerEdge™ R630 Rack Servers)

same memory and processor speed
same number of disk drives in each SMW
same capacity disk drives in each SMW
same capacity disk in the matching drive bays of the two SMWs
(if using R630 SMWs) RAID controller has the same configuration to present the disks to Linux: disks in slot 0 through slot 3 as a virtual disk with RAID5 and the disk on slot 4 not in RAID configuration

Lustre

- This migration process does not include a tested procedure for preserving a DAL file system during migration, though it is expected to be possible. That will be the responsibility of each customer site.
- External Lustre file systems should not require reformatting.
- Direct-attached Lustre (DAL) file systems should not require reformatting.
- DAL LMT (Lustre Monitoring Tool) database will not be migrated.

DataWarp

- If this site has not reformatted/over-provisioned Intel P3608 SSD cards as directed in FN6121a *Datawarp - Performance Issues*, then these Intel P3608 SSD cards must be reformatted. The migration procedures reference the necessary procedure in *XC™ Series DataWarp™ Installation and Administration Guide* (S-2564).
- DataWarp Fusion IO SSDs that are ioMemory3 (for example,SX300) are supported in the CLE 6.0.UP03 / SMW 8.0.UP03 release, but no other models from Fusion IO are supported. The SLES 12 version of SanDisk/Fusion driver (VSL4.2.5) requires firmware version 8.9.5. Sites may need to update (flash) the driver firmware to 8.9.5. However, once updated, the firmware cannot be reverted to the previous version. DO NOT UPDATE FIRMWARE NOW. That will be done later in the migration process.

  CAUTION: Once updated, the firmware revision cannot be reverted to the previous version, so the SSDs will NOT be usable in a CLE 5.2 / SMW 7.2 system.
- DataWarp SanDisk Fusion ioScale2 SSD PCIe boards are no longer supported with CLE 6.0 / SMW 8.0.

Workload managers

Workload manager (WLM) logs are not preserved. If log preservation is desired, speak with the WLM vendor.

Accounting data

Accounting data is not preserved; however, this migration process includes a step for running final accounting reports for the CLE 5.2 / SMW 7.2 system just prior to system shutdown.

FC/SAS/Ethernet cards

Firmware updates for the FC cards, SAS cards, and Ethernet cards that are used in the SMW and CLE nodes should be current. Cray has not tested whether old firmware works after SLES 12 has been installed.

Network interfaces

This migration process does not include a tested procedure for configuring bonded or VLAN network interfaces for a fresh install of CLE 6.0.UP03 / SMW 8.0.UP03.

Warning about Potential Loss of Important Data

WARNING: This migration process includes a fresh install, and when a fresh install is performed on a system, disks are wiped clean. There is a risk of losing important data.

Sites planning to migrate from CLE 5.2 / SMW 7.2 to CLE 6.0 / SMW 8.0 should be aware of the consequences of a migration process that includes a fresh install. This migration process will wipe SMW internal disks and the boot RAID. To prevent loss of important data, phases 3 and 4 of this process provide procedures and extensive guidance for selecting and archiving data prior to shutdown of the CLE 5.2 / SMW 7.2 system.

Here are some ways that SLES 12 and the CLE 6.0.UP03 / SMW 8.0.UP03 release handle SMW and boot RAID storage differently, which is why a fresh install is necessary in this migration process.

- New LUNS created on boot RAID to hold CLE 6.0 / SMW 8.0 file systems.
- Dell R815 SMW uses software RAID1 on two drives for the operating system.
- Dell R630 SMW uses hardware RAID5 on four drives for the operating system.
- Additional SMW disk used for the Power Management (Postgres) database.
- SLES 12 installed on SMW into new disk partitions (`/`, `swap`, `/boot`).
- SMW, boot, and SDB nodes use LVM volume groups on the boot RAID.

3 Migration Training

With the CLE 6.0 / SMW 8.0 releases, Cray has changed the way software is installed, configured, and managed on XC Series systems. Because the CLE 6.0.UP03 / SMW 8.0.UP03 release, which is based on SLES 12, is so different from the CLE 5.2.UP04 / SMW 7.2.UP04 release, which is based on SLES 11, sites experienced with CLE 5.2.UP04 / SMW 7.2.UP04 and older Cray software releases will need to learn about the new Cray Management System (CMS) for a successful migration to the new release.

The new management system uses a common installation process for SMW and CLE, leverages standard Linux and open source tools, and centralizes configuration, keeping configuration data separate from software images until that data is applied to nodes at boot time or whenever cray-ansible is run. The core elements of this new management system are:

IMPS The Image Management and Provisioning System (IMPS) is responsible for creating and distributing repository content and for prescriptive image creation.

CMF The Configuration Management Framework (CMF) comprises the configuration data (stored in config sets on the SMW), tools to manage and distribute that data (e.g., the configurator and the IMPS Distribution System (IDS)), and software to apply the configuration data to the running image (cray-ansible and Ansible plays).

NIMS The Node Image Mapping Service (NIMS) is responsible for keeping track of which images get booted on which nodes, what additional kernel parameters to pass to nodes at boot time, and which load file to use within a boot image.

To help customer sites learn about the new system software, Cray recommends a combination of resources: publicly available Ansible documentation (books and websites), Cray training, Cray technical publications, and the collection of topics in *Supplemental Information* on page 19 at the end of this publication.

Ansible Documentation

Working with the new Cray management software requires a basic understanding of Ansible and Python. Configuration data is applied in large part through Ansible plays, and sites may wish to write their own Ansible plays to supplement that functionality. Some familiarity with the Python programming language will be helpful because Ansible is based on and extendable by Python, and many new Cray tools are written in Python. Get acquainted with this material first, if possible.

Cray recommends reading at least one of these Ansible books (or an equivalent):

- *Ansible: Up & Running: Automating Configuration Management and Deployment the Easy Way*, by Lorin Hochstein
- *Ansible for DevOps: Server and configuration management for humans*, by Jeff Geerling
- *Ansible Playbook Essentials*, by Gourav Shah
- *Mastering Ansible*, by Jesse Keating

Visit these websites for more information about Ansible:

- *https://www.ansible.com/configuration-management*

- *http://docs.ansible.com/*

Cray Training

Cray offers a four-day training course on Cray System Administration, recommended for both site staff and Cray on-site support staff. Even those who have already taken this course with CLE 5.x and SMW 7.x content need to take it again to learn about the new CLE 6.x and SMW 8.x content.

Cray Technical Publications

All Cray technical publications are available at *http://pubs.cray.com*.

To prepare for this migration, Cray strongly recommends reading the following technical publications:

- *What's New for CLE 6.0 and SMW 8.0 (CLE 6.0.UP03) S-2573*
- *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Virtual SMW (CLE 6.0.UP03) S-2575* or *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Physical SMW (CLE 6.0.UP03) S-2580*
- *XC™ Series Configurator User Guide (CLE 6.0.UP03) S-2560*
- *XC™ Series Cray Ansible Writing Guide (CLE 6.0.UP03) S-2582*
- *XC™ Series esLogin to eLogin Migration Guide* (S-2584)
- *XC™ Series eLogin Installation Guide (CLE 6.0.UP03) S-2566 Rev A*
- *XC™ Series eLogin Administration Guide (CLE 6.0.UP03) S-2570 Rev A*
- *XC™ Series Boot Troubleshooting Guide (CLE 6.0.UP03) S-2565*
- *XC™ Series System Administration Guide (CLE 6.0.UP03) S-2393*
- *XC™ Series Software Installation and Configuration Guide (CLE 6.0.UP03) S-2559* (Note that although the detailed migration guides share most of the content of this publication, there may be other content here that would be useful to reference.)
- *XC™ Series Power Management Administration Guide (CLE 6.0.UP03) S-0043*

If these optional features are or will be used at this site, read these publications also:

- DataWarp:
  - *XC™ Series DataWarp™ Installation and Administration Guide (CLE 6.0.UP03) S-2564*, which supersedes *DataWarp Installation Guide S-2547*
  - *XC™ Series DataWarp™ User Guide (CLE 6.0.UP03) S-2558*
- DVS:

  Although the Cray Data Virtualization Service (DVS) is an integral feature of the CLE 6.0.UP03 release, its use by sites to project an external file system or provide access to DataWarp is optional. For those purposes, familiarity with the DVS guide is necessary.
  - *XC™ Series DVS Administration Guide (CLE 6.0.UP03) S-0005*
  - *XC™ Series GPFS Software Installation Guide (CLE 6.0.UP03) S-2569*
- Lustre: *XC™ Series Lustre® Administration Guide (CLE 6.0.UP03) S-2648*
- Shifter:
  - *XC™ Series Shifter Configuration Guide (CLE 6.0.UP03) S-2572*

- *XC™ Series Shifter User Guide (CLE 6.0.UP03) S-2571*

- Slurm: *XC™ Series Slurm Installation Guide (CLE 6.0.UP03) S-2538*
- SMW HA (high availability):
  - *XC™ Series SMW HA Installation Guide (SLEHA12.SP0.UP03) S-0044*
  - *XC™ Series SMW HA Administration Guide (SLEHA12.SP0.UP03) S-2551*
- SEDC: *XC™ Series System Environment Data Collections (SEDC) Guide (CLE 6.0.UP03) S-2491*

4 Migration Planning

For more information and guidance about any of the following planning activities, contact the district service manager (DSM) for this site.

Plan Hardware

This migration process assumes that sites will not upgrade hardware (for example, upgrading Intel® Xeon Phi™ processors, from KNC to KNL) during the migration because of the risk involved in changing software and hardware at the same time. However, some network connection and certain hardware additions may be necessary, depending on the following considerations:

- Is the SDB node connected to the admin network in the CLE 5.2 / SMW 7.2 system?

  Every XC system should have an Ethernet switch with a network for the SMW, boot, and SDB nodes (an "admin" network). Sites that did not connect the SDB node to this admin network in their CLE 5.2 / SMW 7.2 system must do so for migration to a CLE 6.0 / SMW 8.0 system. CLE 6.0 / SMW 8.0 requires that SDB connection. If the system to be migrated does not have an Ethernet card for the SDB node, this site must obtain one.

- How many nodes are available for use as tier2 nodes?

  New nodes may need to be added to the system if an insufficient number of nodes are available. (See "Plan Tier2 Nodes" below.)

- Which type of migration SMW will be used: virtual or physical?

  - If a virtual migration SMW will be used, no additional hardware is required. Cray field support staff will install a virtual SMW with the CLE 6.0.UP03 / SMW 8.0.UP03 release on a Cray laptop (MacBook Pro® or Dell Latitude™), which will be used as the migration SMW. Because of licensing restrictions, no customer computer is authorized to host the virtual SMW.

    Much of the configuration can be done without being connected to XC hardware. Because of virtual SMW disk size limitations, installing PE software must wait until later in the migration process, when CLE 6.0 / SMW 8.0 software is installed on the original SMW that is connected to XC hardware.

  - If a physical migration SMW will be used, a spare SMW and boot RAID must be obtained. The CLE 6.0 / SMW 8.0 software is installed on this SMW before any of the other preparation work is done.

    As with the virtual migration SMW, much of the configuration can be done without being connected to XC hardware, but unlike the virtual SMW, the physical migration SMW has space for installation of the PE software to an image root during the preparation phase.

- Is the CLE 5.2 / SMW 7.2 system to be migrated an SMW HA system?

  If this system is SMW HA, Cray recommends migrating using a physical migration SMW. In this case, an additional spare SMW that matches the physical migration SMW must be obtained:

  same model (both Dell PowerEdge™ R815 Rack Servers or both Dell PowerEdge™ R630 Rack Servers)

same memory and processor speed
same number of disk drives in each SMW
same capacity disk drives in each SMW
same capacity disk in the matching drive bays of the two SMWs
(if using R630 SMWs) RAID controller has the same configuration to present the disks to Linux: disks in slot 0 through slot 3 as a virtual disk with RAID5 and the disk on slot 4 not in RAID configuration

- Does the system being migrated have Intel® Xeon Phi™ "Knight's Corner" (KNC) nodes?

  The CLE 6.0.UP03 / SMW 8.0.UP03 release does not support Intel® Xeon Phi™ "Knight's Corner" (KNC) nodes. If this system has KNC nodes, Cray recommends one of the following options:

  Table 1. What to do with KNC nodes/blades

Situation	Recommended option
Site will replace KNC blades with a blade type that is supported by both CLE 5.2 / SMW 7.2 and CLE 6.0 / SMW 8.0.	Option 1: **1.** Remove the KNC blades and confirm that HSS routing works correctly with the blades gone. **2.** Begin the migration.
Site will not replace KNC blades at all.	same as Option 1
Site will replace KNC blades with a blade type supported by CLE 6.0 / SMW 8.0 but NOT supported by CLE 5.2 / SMW 7.2, such as Intel® Xeon Phi™ "Knight's Landing" (KNL).	Option 2: **1.** Leave the KNC blades in the system but disable the KNC nodes. **2.** Perform the migration. **3.** Remove the blades and confirm that HSS routing works correctly with the blades gone. **4.** Add the new blades and confirm that HSS routing works correctly with the new blades.

Plan Tier2 Nodes

Tier2 nodes are an important part of Cray Scalable Services, which enables configuration data and software on the SMW and boot node to be made available to the rest of the system and log data from the system to be aggregated on the SMW. Cray Scalable Services depends on having a hierarchy of nodes: the Server of Authority (SMW), tier1 nodes (boot and SDB nodes), tier2 nodes (designated service and repurposed compute nodes), and tier3 nodes (everything else). For more information, see *About Cray Scalable Services* on page 19.

Part of hardware planning is to ensure the correct ratio of tier3 nodes (clients) to tier2 nodes (servers). On a CLE 5.x system, the DSL nodes, which NFS-mounted the sharedroot and then DVS-projected it to compute nodes, are similar but not the same as tier2 nodes. There may be enough DSL nodes from CLE 5.x to be reused as tier2 nodes, but more tier2 nodes may be required. See the tier2 node FAQ below for specific rules on how many tier2 nodes are required to support the number of tier3 nodes and which type of nodes can be used as tier2 nodes.

Q. How many tier2 nodes are needed?	**A.** At least one server must be provided, and for resiliency, two nodes placed on different blades. The recommended ratio of tier2 nodes (servers) to tier3 nodes (clients) is 1 to 400.

Q. Will adding more tier2 nodes help performance?

A. Adding more tier2 nodes does not always yield additional performance and is subject to diminishing returns.

Q. What kind of node can be used as a tier2 node?

A. Use these:

- OPTIMAL: dedicated repurposed compute nodes (RCN)
- dedicated service nodes
- nodes with uniform light to moderate load
- nodes with relatively homogeneous single core speeds to reduce resource contention disparity during periods of partial availability

AVOID these (will result in sub-optimal performance):

- nodes with slower individual CPU cores, such as Intel® Xeon Phi™ processors (formerly code named Knights Landing or KNL)
- direct-attached Lustre (DAL) servers
- RSIP (realm-specific IP) servers
- login nodes

Q. Can a tier2 node have more than one role?

A. Small test and development systems (TDS) may use tier2 nodes that have additional roles, but generally, it is better for tier2 nodes to be dedicated.

Q. Where should tier2 nodes be placed?

A. Distribute nodes throughout the system (on different blades) for resiliency in the event of hardware failure.

5 Preparation of Configuration Data and Software Images

To minimize the downtime required to switch an XC system from CLE 5.2 / SMW 7.2 to CLE 6.0 / SMW 8.0, configuration data and software images can be prepared ahead of time. This phase of the migration process requires access to a migration SMW with CLE 6.0 / SMW 8.0 installed, which is used to stage the changes for configuration and image management. Many of the steps are the same whether the migration SMW is virtual or physical.

These steps are only a preview of the tasks that will be required in this phase of the migration. They are not intended to be performed as written. Task details are provided in *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Virtual SMW* (S-2575) and *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Physical SMW* (S-2580).

1. Perform these procedures on the system running CLE 5.2 / SMW 7.2:
   a. Ensure either access to the physical keyboard, mouse, and monitor of the original SMW (running CLE 5.2 / SMW 7.2) or connection over iDRAC (if that SMW is not physically present) before trying to perform the migration.
   b. Start a typescript file to capture commands and output.
   c. Extract configuration information from the current system (running CLE 5.2 / SMW 7.2).
2. Perform these procedures on the migration SMW, which is running CLE 6.0 / SMW 8.0:
   a. Read man pages, when an SMW running CLE 6.0 / SMW 8.0 is available, to increase familiarity with the new CLE 6.0 / SMW 8.0 commands.
   b. Transfer configuration information from the currently running (CLE 5.2 / SMW 7.2) system into configuration worksheets on the migration SMW.

      This step takes the configuration information extracted from the CLE 5.2 / SMW 7.2 system and enters it into configuration worksheets on the migration SMW. See *Cray XC System Configuration* on page 21. Many of the configuration worksheets include *node group* settings/variables, which are similar to but more powerful than the node class specialization of releases prior to CLE 6.0 / SMW 8.0. See *About Node Groups* on page 24.
   c. Load and validate configuration worksheets on the migration SMW to validate the worksheets and the consistency of the configuration data.
   d. Update non-config-set configuration files.
   e. Choose which image recipes to build.

      In addition to the basic set of images that are needed to boot CLE—admin, service, login, and compute—additional images may be needed for DAL (direct-attached Lustre), DataWarp with Fusion I/O SSDs, and Netroot (see *Where to Place the Root File System—tmpfs versus Netroot* on page 23). Sites may need to extend image recipes with workload manager (WLM) content, site-specific RPMs, non-RPM content, or to run certain commands in a chrooted context as the recipe is built into an image root.
   f. Build image roots and boot images.

g. Assign kernel parameters to nodes.

h. Check NIMS data.

i. Identify and port site-local scripts (done on current system + migration SMW).

j. (If a virtual migration SMW is used) Prepare tar archives on virtual migration SMW.

k. (If a physical migration SMW is used) Install Cray Programming Environment (PE) Software.

6 Preservation of Other Data Prior to Final Shutdown

These steps should be performed just before shutting down the CLE 5.2 / SMW 7.2 system for the last time, so that the most current site accounting, operational, and user data is preserved. Cray recommends saving this data separately from the configuration and image data captured and archived in the previous stage of this migration process.

1. Run final accounting reports.

   If process accounting, SAR (system activity reporter), or RUR (resource utilization reporting) data has been generated on the CLE 5.2 / SMW 7.2 system that needs to be used for accounting or billing based on application job usage, run those reports just before the XC system is shut down.

2. Save operational data.

   Save CLE 5.2 / SMW 7.2 operational data, such as boot automation files, tuning settings in HSS configuration files, crontabs, error logs, and other hardware status information.

   - mail configuration files on SMW
   - old CLE dumps
   - workload manager (WLM) files in `/var` on the WLM server and MOM nodes
   - SEC logs on SMW
   - SEDC files on SMW
   - SMW firewall settings
   - SMW logs and CLE logs for recent boot sessions

3. Save site user data.

   Save any CLE 5.2 / SMW 7.2 site user data, such as home directories for Linux accounts, workload manager logs, and user DataWarp files on the SSDs.

4. Drain WLM queues before shutting down.

7 Shutdown and Switch

The final phase of the migration process is to shut down the currently running system and make the switch to the new release. This is another point at which the migration process diverges, depending on whether the migration SMW is virtual or physical.

virtual SMW	If a virtual migration SMW is used, then the configuration data and images must be moved from the virtual migration SMW to the original (physical) SMW that is connected to the XC hardware. This necessitates a fresh install of all software on the original SMW.
physical SMW + boot RAID	If a physical migration SMW and boot RAID are used, then the original SMW and boot RAID must be disconnected from the XC hardware and the migration SMW and boot RAID must be connected to the XC hardware.

Virtual SMW migration: fresh install on original SMW and boot RAID

These steps are only a preview of the tasks that will be required in this phase of the migration. They are not intended to be performed as written. Task details are provided in *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Virtual SMW* (S-2575).

1. Shut down the CLE system.
2. Wipe data on the boot RAID and SMW disks while installing the new SLES 12, SMW, and CLE software on the original (physical) SMW.
3. Configure SMW for CLE system hardware (includes XC system hardware discovery and updating firmware on components).
4. Complete CLE configuration with hardware connected.
5. Install Cray Programming Environment (PE) software to image root.
6. Complete first boot of CLE nodes with new software.
7. Complete post-boot configuration of config services.
8. Configure other CLE 6.0 / SMW 8.0 features and services and install additional software (including SMW HA).
9. Restore any CLE 5.2 / SMW 7.2 operational data (files, database exports, site user data, and site-local scripts).

Physical SMW migration: switch from original SMW and boot RAID to migration SMW and boot RAID

As in the virtual SMW scenario, these steps are only a preview of the tasks that will be required in this phase of the migration. They are not intended to be performed as written. Task details are provided in *XC™ Series CLE 5.2 to CLE 6.0 Software Migration using a Physical SMW* (S-2580).

1. Shut down the CLE system.
2. (SMW HA only) Put the SMW HA cluster in maintenance mode.
3. Switch cabling to the physical migration SMW and boot RAID.
4. Discover XC system hardware and update firmware on components.

5. Complete CLE configuration with hardware connected.
6. Complete first boot of CLE nodes with new software.
7. Configure other CLE 6.0 / SMW 8.0 features and services and install additional software (including SMW HA).
8. Restore any CLE 5.2 / SMW 7.2 operational data (files, database exports, site user data, and site-local scripts).

8 Supplemental Information

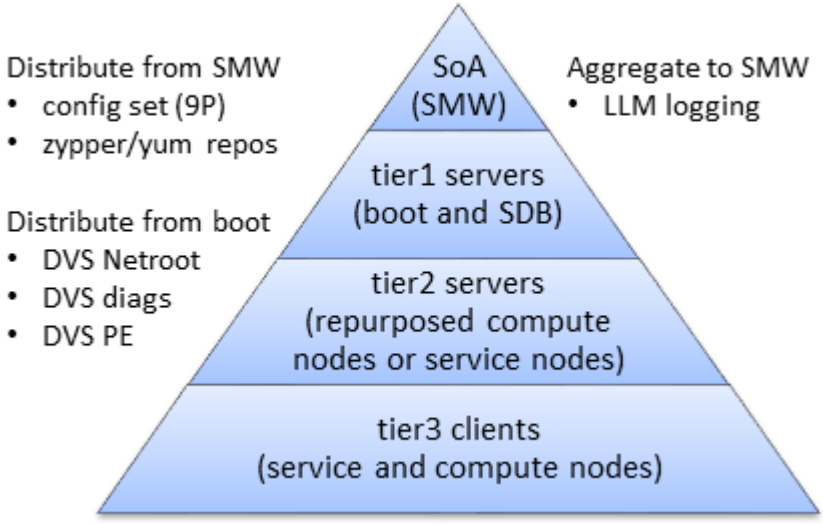
This collection of topics is provided as background information about some aspects of the new Cray management system software referenced in this migration overview.

8.1 About Cray Scalable Services

Cray Scalable Services is an essential part of the Cray Management System that is used to both distribute and aggregate information. Within Cray Scalable Services, nodes are designated as SoA (server of authority), tier1, tier2, or tier3. A node can be a member of only one of these groups. Tier1 nodes are clients of the SoA and servers for tier2 nodes. Tier2 nodes are clients of tier1 nodes and servers for tier3 nodes. Tier3 nodes are clients of tier2 nodes. Configuration of nodes as SoA, tier1, and tier2 is defined in the cray_scalable_services configuration service, which must be configured properly for the system to function.

As indicated in this figure, the SMW is the designated SoA in Cray XC systems. The boot and SDB nodes are designated tier1 nodes, and they must have direct network connectivity to the SMW via Ethernet. Typically, tier2 nodes are service nodes or repurposed compute nodes that have no other duties beyond being part of the Scalable Services. All other nodes are tier3 nodes.

Figure 1. Cray Scalable Services



This table shows what gets distributed or aggregated using Cray Scalable Services.

from SMW to rest of system	• config set data is shared using a 9P file system and DIOD (distributed I/O daemon) • zypper software repositories can by used from any node with the Live Update feature (http forwarding from the SMW through the tiers)
from boot node to rest of system	• PE (Programming Environment) image root • diag (online diagnostics) image root • Netroot image roots[1]
from rest of system to SMW	• Lightweight Logging Manager (LLM) logging

Here is an example of how Scalable Services works with Live Updates to distribute software out to nodes. Any tier3 node can run zypper to access the repositories on the SMW because it has an entry in `/etc/zypp/repos.d/liveupdates.repo` that points to the tier2 nodes by means of a baseurl, which uses http protocol listing all of the tier2 nodes. The tier2 nodes, in turn, have an entry in `/etc/zypp/repos.d/liveupdates.repo` that lists at least one tier1 node. All tier1 nodes have an entry in `/etc/zypp/repos.d/liveupdates.repo` that lists the SMW.

Services that Depend on Cray Scalable Services

It is important to configure Cray Scalable Services correctly. The following features and services use data from the cray_scalable_services configuration service, and may they not be functional if cray_scalable_services is configured incorrectly.

Node Image Mapping Service (NIMS) plugin	Uses cray_scalable_services data to determine tier1 servers and adds the tier1 kernel command line parameter to each tier1 server.

[1] Netroot is a mechanism that enables nodes booted with a minimal, local in-memory file system to execute within the context of a larger, full-featured root file system which available to the node via a network mount.

IMPS Distribution Service (IDS)	Uses cray_scalable_services data to set the ids kernel command line parameter to the node's parent, from whom it will receive config set data.
DVS Ansible configuration	Uses cray_scalable_services data to determine which nodes should serve DVS file systems. This will also impact Netroot functionality, which uses DVS.
CLE liveupdates functionality	Configured using cray_scalable_services data to determine the parent each node should contact en route to the package repos stored on the SMW.
LLM Ansible configuration	Uses cray_scalable_services data to determine the next server to which a node should send its log data, which depends on the node's tier.
NFS Ansible configuration	Uses cray_scalable_services data to determine which nodes should act as clients and servers.
IP forwarding Ansible configuration	Uses cray_scalable_services data to enable IP forwarding and configure servers' routes depending on their tier.

8.2 Cray XC System Configuration

To configure Cray XC systems and manage configuration content, system administrators use the Cray configuration management framework (CMF). The CMF comprises configuration data, the tools to manage and distribute that data, and software to apply the configuration data to the running image at boot time. Its major components include configuration service packages, config sets, the IMPS distribution service (IDS), the configurator, cray-ansible, and Ansible.

Configuration Starts with Configuration Service Packages

Configuration content (data and software) is installed as configuration service packages on the management node of Cray XC systems (in `/opt/cray/imps_config/<service package>/default/configurator` by default). Each service package delivers configuration content for one or more system services. The contents of each service package reside in the following subdirectories:

`ansible`	Drop zone for Cray-provided Ansible play content.
`callbacks`	Pre- and post-configuration scripts.
`dist`	Drop zone for other Cray-provided content, such as static files required for the configuration of a service.
`template`	Configuration templates that define the configuration settings to be set and provide some default values. These templates are never modified by administrators or other users.

Configuration service packages are installed for system upgrades and updates as well as for initial installation.

Configuration Information is Stored in Config Sets

Administrators use the `cfgset` command to manage configuration information. It takes configuration content delivered in service packages and invokes the *configurator* tool to combine that content with site-specific configuration content gathered from administrators either interactively or through bulk import. The results are used by `cfgset` to create a configuration set or *config set*. A config set is a central repository that stores all configuration information necessary to operate the system. Config sets reside on the management node (e.g., the

SMW) in `/var/opt/cray/imps/config/sets` by default. The contents of each config set reside in the following subdirectories:

`ansible` Drop zone for local site-provided Ansible play content to be distributed with the config set. When the config set is created, `cfgset` copies Ansible content from service packages to this location. Whenever the config set is updated, `cfgset` copies Ansible content from service packages again, overwriting the previous service-package Ansible content and leaving the site-provided content unchanged.

`changelog` YAML change logs from previous sessions with the configurator.

`config` Configuration templates containing configuration information. When the config set is created, the configurator copies service package templates to this location. Administrators can modify the content of these templates using `cfgset` and the configurator. Whenever the config set is updated, the configurator merges service package templates with the templates in this location.

`dist` Drop zone for other site-provided content, such as static files required for the configuration of a service. When the config set is created, `cfgset` copies dist content from service packages to this location. Whenever the config set is updated, `cfgset` copies dist content from service packages again, overwriting the previous service-package dist content and leaving the site-provided content unchanged.

`files` Files necessary for system configuration that are generated by configuration callback scripts or manually and distributed with the config set (e.g., `/etc/hosts`).

`worksheets` Configuration worksheets generated by the configurator using data stored in the configuration templates in the `config` subdirectory of the config set. Administrators copy these worksheets to a location outside the config set, edit them with site-specific configuration data, and then import them to create a new config set or update an existing one.

An administrator may create multiple config sets to support partitions or alternate configurations. Typically a config set of type `cle` is created for each partition to store partition- and CLE-specific content, and another config set of type `global` is created to store management node and global configuration data.

IDS Distributes Config Sets to Nodes

IDS, a read-only network share of content from the management node to the rest of the system, distributes config sets to every node in the system. All config sets are shared throughout the system, but only one `cle` config set is active on a given node at a time (in addition to an active `global` config set, which is applied to the entire system). Currently, IDS leverages the 9P network file system and the Linux automounter facility as its distribution mechanism; however, the content and use of the config sets is independent of the distribution mechanism.

Ansible Plays Apply Configuration during System Boot

Prior to booting the system, each node will have an image, the `global` config set, and the `cle` config set. When the system boots, each node boots an unconfigured software image. Then Ansible plays, which can be located in both the image and the config set (config set is the preferred location for site-supplied Ansible plays), apply configuration to that image, bringing up the services pertinent to each node.

Administrators Configure/Reconfigure the System on an Ongoing Basis

Configuration happens at times other than initial installation. New configuration service packages can be installed during system upgrades and updates, sites can decide to enable a new service or change the configuration of an existing service, and so forth. In all of these scenarios, an administrator uses the `cfgset` command to manage

config sets and the `cray-ansible` script to apply any configuration changes. The `cfgset` command and its associated subcommands and options enable administrators to perform a variety of operations on config sets in addition to create and update, such as search, diff, list, show, validate, push, and remove. See the `cfgset` man page for a description of its subcommands and options and some examples of each.

8.3 Where to Place the Root File System—tmpfs versus Netroot

The Cray XC™ Series root file system for nodes can either reside in RAM (tmpfs) or be mounted from a network source (Netroot), depending on the type of node. The boot and SDB nodes, all other service nodes (except login nodes), and all DAL (direct-attached Lustre) nodes must use tmpfs. Compute nodes and login nodes may use either tmpfs or Netroot. Use the information provided here to decide whether to use Netroot for some or all compute and login nodes at this site.

About Netroot and Dynamic Shared Objects and Libraries (DSL)

In releases prior to CLE 6.0 / SMW 8.0, the dynamic shared objects and libraries (DSL) feature was optional. It was necessary for many sites because it enabled both dynamic shared libraries and large network-based images, which were needed for systems with NVIDIA GPUs and for most production workloads.

In the current release, DSL is supported by default. Note, however, that the DSL feature no longer includes provision for large network-based images. That capability is now provided by Netroot.

- Sites that require large network-based images and additional storage should use Netroot.
- Sites using NVIDIA GPUs must use Netroot.

Comparison of tmpfs and Netroot

tmpfs The default location of the root file system on Cray XC™ Series systems is tmpfs, a type of memory-resident file system or RAM disk.

tmpfs has these characteristics and limitations:

- always used for service nodes (except login nodes) and DAL (direct-attached Lustre) nodes
- efficient and fast root file system access
- large memory footprint
- file system content needs to be restricted to reduce memory footprint
- typically used when minimal commands and libraries required
- works well for compute nodes with well defined workloads and for service nodes that are used primarily for internal services

Netroot Netroot is an alternative approach that mounts the root file system from a network source. It is used only for compute and login nodes. It uses overlayfs to layer tmpfs on top of a read-only network file system.

Due to the reliance on overlayfs, the decision to use Netroot should include consideration of the characteristics and limitations of overlayfs in addition to those of Netroot listed here.

Netroot has these characteristics and limitations:

- used only for compute and login nodes, never for service nodes (except login nodes)
- slower root file system access
- increased node boot time
- minimized memory footprint (mounted from network, so requires less disk space)
- no restriction on file system content
- typically used when a robust set of commands and libraries required (Netroot enables large network-based images, formerly enabled through the DSL feature)
- works well for compute nodes with diverse workloads and for compute nodes with a high memory footprint
- always used for GPUs

This comparison of tmpfs and Netroot memory footprints is based on a fresh install with nothing extra added. These numbers could be larger or smaller for a site depending on whether the Cray image recipes for tmpfs and Netroot have been extended (by adding necessary RPMs) or reduced (by removing unnecessary RPMs).

Table 2. Comparison of tmpfs and Netroot Memory Footprints

Image Type	Memory Consumption	Number of RPMs
Admin image root - tmpfs	1400 MB	600
Service image root – tmpfs	1700 MB	670
Login image root – tmpfs	3600 MB	1100
Compute image root – tmpfs	1500 MB	745
Login image root – Netroot	125 MB	2500
Compute image root – Netroot	150 MB	2380

8.4 About Node Groups

The Cray Node Groups service (cray_node_groups) enables administrators to define and manage logical groupings of system nodes. Nodes can be grouped arbitrarily, though typically they are grouped by software functionality or hardware characteristics, such as login, compute, service, DVS servers, and RSIP servers.

Node groups that have been defined in a config set can be referenced by name within all CLE services in that config set, thereby eliminating the need to specify groups of nodes (often the same ones) for each service individually and greatly streamlining service configuration. Node groups are used in many Cray-provided Ansible configuration playbooks and roles and can be also used in site-local Ansible plays. Node groups are similar to but more powerful than the class specialization feature of releases prior to CLE 6.0. For example, a node can be a member of more than one node group but could belong to only one class.

Sites are encouraged to define their own node groups and specify their members. Administrators can define and manage node groups using any of these methods:

- Edit and upload the node groups configuration worksheet (`cray_node_groups_worksheet.yaml`).

- Use the `cfgset` command to view and modify node groups interactively with the configurator.
- Edit the node groups configuration template (`cray_node_groups_config.yaml`) directly. Use `cfgset` to update the config set afterwards so that pre- and post-configuration scripts are run.

After using any of these methods, remember to validate the config set.

Characteristics of Node Groups

- Node group membership is not exclusive, that is, a node may be a member of more than one node group.
- Node group membership is specified as a list of cnames. However, if the SMW is part of a node group, it is specified with the output of the `hostid` command. Also, host names can be used for eLogin nodes that are to be included in node groups.
- All compute nodes and/or all service nodes can be added as node group members by including the keywords “platform:compute” and/or “platform:service” in a node group.
- Any CLE configuration service is able to reference any defined node group by name.
- The Configuration Management Framework (CMF) exposes node group membership of the current node through the local system "facts" provided by the Ansible runtime environment. This means that each node knows what node groups it belongs to, and that knowledge can be used in Cray and site-local Ansible playbooks.

Default Node Groups

Default node groups are groups of nodes that

- are likely to be customized and used by many sites
- support useful default values for many of the migrated services

Several of the default node groups require customization by a site to provide the appropriate node membership information. This table lists the Cray default groups and indicates which ones require site customization.

Table 3. cray_node_groups

Default Node Group	Requires Customization?	Notes
compute_nodes	No	Defines all compute nodes for the given partition. The list of nodes is determined at runtime.
service_nodes	No	Defines all service nodes for the given partition. The list of nodes is determined at runtime.
smw_nodes	Yes	Add the output of the `hostid` command for the SMW. For an SMW HA system, add the host ID of the second SMW also.
boot_nodes	Yes	Add the cname of the boot node. If there is a failover boot node, add its cname also.
sdb_nodes	Yes	Add the cname of the SDB node. If there is a failover SDB node, add its cname also.
login_nodes	Yes	Add the names of internal login nodes on the system.

Default Node Group	Requires Customization?	Notes
all_nodes	Maybe	Defines all compute nodes and service nodes on the system. Add external nodes (e.g., eLogin nodes), as needed.
tier2_nodes	Yes	Add the cnames of nodes that will be used as tier2 servers in the cray_scalable_services configuration.

Why is there no "tier1_nodes" default node group? Cray provides a default tier2_nodes node group to support defaults in the cray_simple_shares service. Cray does not provide a tier1_nodes node group because no default data in any service requires it. Because it is likely that tier1 nodes will consist of only the boot node and the SDB node, for which node groups already exist, Cray recommends using those groups to populate the cray_scalable_services tier1_groups setting rather than defining a tier1_nodes group.

About eLogin nodes. To add eLogin nodes to node groups, use their 'hostname' values instead of cnames, because unlike CLE nodes, eLogin nodes do not have cname identifiers. If eLogin nodes are intended to receive configuration settings associated with the all_nodes group, add them to that group, or create a new group for eLogin nodes only (elogin_nodes), and then change the appropriate settings in other configuration services to include both all_nodes and elogin_nodes.

Additional Platform Keywords

Cray uses these two platform keywords to create default node groups that contain all compute or all service nodes.

```
platform:compute
platform:service
```

Sites that need finer-grained groupings can use these additional platform keywords to create custom node groups that contain all compute or service nodes with a particular core type.

```
platform:compute-XXNN
platform:service-XXNN
```

For XXNN, substitute a four-character processor/core designation, such as KL64 or KL68, which designate the two Intel® Xeon Phi™ processors (Knights Landing) with different core counts.

Table 4. Cray Supported Intel Processor/Core (XXNN) Designations

Processor (XX)	Core (NN)	Intel Code Name
BW	12, 14, 16, 18, 20, 22, 24, 28, 32, 36, 40, 44	Broadwell
HW	04, 06, 08, 10, 12, 14, 16, 18, 20, 24, 28, 32, 36	Haswell
IV	02, 04, 06, 08, 10, 12, 16, 20, 24	Ivy Bridge
KL	60, 64, 66, 68, 72	Knights Landing
SB	04, 06, 08, 12, 16	Sandy Bridge