ClusterStor™ System Update Bundle Installation Guide 3.x.x

(3.0.0.SU015)

S-2596

Contents

1 About the ClusterStor System Update Bundle Installation Guide 3.X.X

The *ClusterStor™ System Update Bundle Installation Guide (3.0.0.SU015) S-2596* provides installation instructions for the system update (SU) bundle 3.0.0.SU015 software release.

Scope and Audience

This publication is intended for use by Cray service technicians who are trained in ClusterStor system administration.

Typographic Conventions

`Monospace`	Indicates program code, reserved words, library functions, command-line prompts, screen output, file/path names, and other software constructs.
`Monospaced Bold`	Indicates commands that must be entered on a command line or in response to an interactive prompt.
`Oblique` or *`Italics`*	Indicates user-supplied values in commands or syntax definitions.
Proportional Bold	Indicates a **GUI Window**, **GUI element**, cascading menu (**Ctrl**→**Alt**→**Delete**), or key strokes (press **Enter**).
\ (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).

Other Conventions

Sample commands and command output used throughout this publication are shown with a generic filesystem name of **cls12345**.

Trademarks

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, Urika-GX, and YARCDATA. The following are trademarks of Cray Inc.: APPRENTICE2, CHAPEL, CLUSTER CONNECT, ClusterStor, CRAYDOC, CRAYPAT, CRAYPORT, DATAWARP, ECOPHLEX, LIBSCI, NODEKARE. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

2 Introduction

This system update (SU) contains a self-extracting archive: `su_csl-3.0.0-015.91.zip`. It contains packages that fix system-wide problems (including firmware and Lustre).

System Update Script

The archive includes the system update script: `system-update_csl-3.0.0-015.91.sh`, which is used to install the update.

Script Command Line Options

`prepare`: Prepare repositories and the system before running the update.

`install`: Apply update to management and appliance nodes and images.

`post-install`: Perform actions which must occur after the system update install step.

`verify`: Verify that system updates were installed with no remaining repository actions.

`check`: not used.

Command Line Conventions Used in this Document

The following conventions are used for command line prompts.

`[MDS]$` – command entered from MDS node as admin.

`[MDS]#` – command entered from MDS node as root.

`[MGMT0]$` – command entered from primary management node as admin.

`[MGMT1]$` – command entered from secondary management node as admin.

`[MGMT0]#` – command entered from primary management node as root.

`[MGMT1]#` – command entered from secondary management node as root.

3 Prepare to Install the System Update Bundle

Prerequisites

The following prerequisites pertain to both the preparation and installation procedures for ClusterStor system update 3.0.0.SU015.

System Access Requirements

Parts of the installation procedure described in this document require the technician to have root (super user) access. Contact Cray Support for root access, if necessary.

Service Interruption Level

Interrupt: The installation procedure requires taking the Lustre file system offline.

Required Files

`su_csl-3.0.0-015.91.zip`

Estimated Time Required

The estimated time to complete the system update installation procedure is approximately 4 hours. GEM GOBI firmware update is estimated at approximately 8 hours.

Firmware Packages

Listed below are the latest firmware versions available for ClusterStor. For more information, refer to the firmware change history in the release notes.

Latest USM Firmware Package	Instructions Reference
stx_usm_sbb-onestor-4.1.16a.3.0-1.0.18ae96.X86_64.rpm	For additional information about updating USM firmware on 3.0.0 systems, please contact Cray Support.
Note that the USM firmware update procedure is required only for ESUs and the optional AMMU EBOD, if used in the system.	

Latest GEM GOBI Firmware Package	Instructions Reference
stx_package-3.0-1.33.4.x86_64.RPM	For additional information about updating GEM GOBI firmware on 3.0.0 systems, please contact Cray Support.
Note that the GEM GOBI firmware update procedure is a major undertaking that requires 8 hours or more. Because of difficulties that may be encountered when performing GOBI firmware updates, Cray strongly suggests that customers do not perform the procedure without the supervision and assistance of Cray Support.	

See section *Verify and Update the Firmware* on page 17 for update instructions.

About this task

Perform these steps before installing 3.0.0.SU015.

Procedure

1. If the ClusterStor system has any non-standard tuning or other customization, note them prior to installing the system update.

   Default settings are applied if a parent software package includes specific site customizations. These customizations must be manually re-applied after this update.

2. Log in to MGMT0 and sudo to root.

   ```
   [MGMT0]$ sudo su -
   ```

3. From the primary (active) management node, verify the system is in daily mode.

   ```
   [MGMT0#] cscli cluster_mode --status
   ```

   If there is no output from the command, or anything other than "Daily" mode displays, do not proceed with update. Open a case with information about how to work around the problem or proceed. See CAST-15604.

4. SSH to MGMT1.

5. Create the directory `~admin/300.SU015` on MGMT1:

   ```
   [MGMT1]# mkdir ~admin/300.SU015
   ```

6. Change directory to `~admin/300.SU015`:

   ```
   [MGMT1]# cd ~admin/300.SU015
   ```

7. Copy or download `su_csl-3.0.0-015.91.zip` and associated release notes into the `300.SU015` directory.

8. Extract the ZIP archive.

   If errors occur during the extraction process, copy or download the ZIP file again.

   ```
   [MGMT1]# unzip su_csl-3.0.0-015.91.zip
   Archive: su_csl-3.0.0-015.91.zip
   inflating: system-update_csl-3.0.0-015.91.sh
   inflating: SHA1SUM-bundles.txt
   ```

9. Use the `SHA1SUM-bundles.txt` file to validate that the contents of the ZIP extracted successfully and are not corrupt.

   ```
   [MGMT1]# sha1sum -c SHA1SUM-bundles.txt
   system-update_csl-3.0.0-015.91.sh : OK
   ```

10. Stop client I/O and unmount all Lustre clients. This can vary by site configuration. Basic configurations may simply be client unmount commands as root:

```
[Client]# umount lustre_mount_point
```

11. Exit to return to MGMT0.

12. If the system is NXD enabled, perform the following steps. Otherwise, go to step *13* on page 7.

a. Stop NXD caching on the system.

```
[MGMT0]# cscli nxd disable
```

This command initiates a process to flush the NXD cache to the main array. If the cache is completely full, it may take more than 40 minutes to flush all cached data out to drives. Note that there is no output from the `cscli nxd disable` command.

b. Monitor the NXD caching state by running the `cscli nxd list` command repeatedly and checking the output. Note that while the cache is being flushed, the value displayed in the **Caching State field** (of the `cscli nxd list` command output) for the OSS nodes and OSTs is `disabling`. The displayed values change to `disabled` once the cache is completely flushed.

c. Wait until the value displayed in the **Caching State** field is `disabled` for all OSS nodes and OSTs before proceeding to the next step.

```
[root@cls12345n000 ~]# cscli nxd list
---------------------------------------------------------------------------------------
Host Cache Caching Total Cache Cache Cache Bypass
Group State Cache Size Block Window IO
Size In Use Size Size Size
----------------------------------------------------------------------------------
-----
cls12345n004 nxd_cache_0 disabled 1.406 TB 699.875 MB 8(4 KiB) 128(64 KiB) 64(32 KiB)
cls12345n005 nxd_cache_1 disabled 1.406 TB 745.500 MB 8(4 KiB) 128(64 KiB) 64(32 KiB)
---------------------------------------------------------------------------------------
```

13. After all clients have been identified and unmounted, stop the Lustre file system:

```
[MGMT0]# cscli unmount -f filesystem_name
unmount: No resources found on nodes cls12345n[000-001] for "cls12345" file
system
unmount: stopping cls12345 on cls12345n[002-003]...
unmount: stopping cls12345 on cls12345n[004-005]...
unmount: cls12345 is stopped on cls12345n[002-003]!
unmount: cls12345 is stopped on cls12345n[004-005]!
unmount: MGS is stopping...
unmount: MGS is stopped!
unmount: File system cls12345 is unmounted.
```

If any clients are still mounted, add the arguments `--evict --force` to the end of the `cscli unmount` command to force eviction when the file system stops:

```
[MGMT0]# cscli unmount -f filesystem_name [--evict --force]
```

14. Verify that the Lustre file system has stopped on all nodes:

```
[MGMT0]# cscli fs_info
-----------------------------------------------------------------------
Information about "cls12345" file system:
-------------------------------------------------------------------------------
Node         Node  Targets  Failover      Devices
             type           partner
-------------------------------------------------------------------------------
cls12345n002 mgs   0 / 0    cls12345n003
```

```
cls12345n003 mds   0 / 1    cls12345n002                   /dev/md66
cls12345n004 oss   0 / 4    cls12345n005  /dev/md0, /dev/md2, /dev/md4, /dev/md6
cls12345n005 oss   0 / 4    cls12345n004  /dev/md1, /dev/md3, /dev/md5, /dev/md7
```

In the output, under the heading “Targets,” look for “0” in the first character position, for example, 0 / 4 or 0 / 0. If all values are “0,” then Lustre has been stopped. If repeated attempts continue to show non-"0" targets, try:

```
[MGMT0]# pdsh -a mount -t lustre
```

"No output" implies that Lustre has stopped.

Perform the system update installation procedure.

4 Install ClusterStor System Update Bundle 3.0.0.SU015

Prerequisites

System Access Requirements

Root privilege is required to perform this procedure. Contact Cray Support for root privilege if necessary.

Service Interruption Level

This procedure requires that the Lustre file system be offline.

Estimated Service Time

The estimated length of time to complete the system update procedure is approximately 4 hours. If firmware leveling has been recommended, refer to *Verify and Update the Firmware* on page 17. For additional service time requirements related to updating GEM GOBI firmware and USM firmware on 3.0.0 systems, please contact Cray Support.

Preparation

Complete *Prepare to Install the System Update Bundle* on page 5 before starting this procedure.

About this task

Perform the following steps to install the system update bundle for version 3.0.0-SU015.

Procedure

1. If not already in an SSH session with the primary MGMT node, connect to the primary MGMT node via SSH and change to root user:

   ```
   remote $ ssh -l admin MGMT0
   [MGMT0]$ sudo su -
   ```

2. Identify the MGMT node hosting NFS services and check the location of the `md67` resource group:

   ```
   [MGMT0]# crm_mon -1r
   ```

3. If the `md67` resource group is on MGMT0, fail back the resource to MGMT1:

   ```
   [MGMT0]# cscli failback -n MGMT1
   ```

   When `crm_mon -1r` shows all resources started, with `md67` on MGMT1, continue.

4. Log in to the MGMT NFS server node (`MGMT1`) via SSH:

```
[MGMT0]# ssh MGMT1
```

5. Change directory to `300.SU015`:

   ```
   [MGMT1]# cd ~admin/300.SU015
   ```

6. Start a screen session. A log will be kept in `~admin/300.SU015/screenlog.0`:

   ```
   [MGMT1]# screen -L -S update
   ```

   If the SSH session is interrupted, the update script should continue to run in the screen session. To reconnect:

   a. Log into the NFS host node via SSH.

   b. Run `screen -ls` to confirm the screen is still active.

   c. Run `screen -L -RDS update` to reconnect to the screen session.

7. Make the script executable:

   ```
   [MGMT1]# chmod +x system-update_csl-3.0.0-015.91.sh
   ```

8. From the primary (active) management node, verify the system is in daily mode.

   ```
   [MGMT0#] cscli cluster_mode --status
   ```

   If there is no output from the command, or anything other than "Daily" mode displays, do not proceed with update. Open case with information about how to work around the problem or proceed. See CAST-15604.

9. Prepare the system to install the bundle:

   ```
   [MGMT1]# ./system-update_csl-3.0.0-015.91.sh prepare
   Extracting packages from bundle. All debug output
   will be saved to /var/log/system-update_cs-prepare-yyyymmddhhmmss.log
   .............................................................
   prepare: SUCCESS
   ```

10. Install the bundle:

    IMPORTANT: If non-ClusterStor RPMs are installed on the system, for example, System Snapshot Analyzer (SSA) software, these may need to be re-installed following the update.

    The time required for the installation step depends in part on the content, and the delta from the current version to the version being installed. This step can take at least 30 minutes and sometimes more than 1 or 2 hours.

    ```
    [MGMT1]# ./system-update_csl-3.0.0-015.91.sh install
    MGMT nodes will now be updated with system-update_csl-3.0.0-015.91.
    All debug output will be saved to /var/log/system-update_cs-install-
    yyyymmddhhmmss.log
    Cleaning metadata........
    Checking if all MGMT resources needed for SU installation are started
    All required resources are up and running, installation can proceed
    Waiting for HA to become stable....DONE
    Finishing remaining transactions from previous installs....
    Following packages are no longer required and will be removed
     . . .
    install: SUCCESS
    ```

Note that some warnings are normal. As long as the installation completes and the subsequent verification steps succeed, the update has been applied successfully. Examples of warnings include:

```
[warn] worker http://localhost:8022/ already used by another worker
WARNING: /lib/modules/2.6.32-220.7.1.el6.lustre.4116.x86_64/kernel/drivers/
infiniband/
ulp/srp/ib_srp.ko needs unknown symbol ib_wq
mysqladmin: CREATE DATABASE failed; error: ’Can’t create database ’t0db’;
database exists’
```

11. Exit the screen session.

```
[MGMT1]# exit
[screen is terminating]
```

12. Exit the SSH session to return to the primary management node.

```
[MGMT1]# exit
```

13. Reboot the system where the install was just performed:

a. Power off all the diskless nodes:

```
[MGMT0]# pm -0 $(nodeattr -s diskless)
```

b. Verify that the nodes are off:

```
[MGMT0]# pm -q $(nodeattr -s diskless)
```

c. Fail over MGMT1:

```
[MGMT0]# cscli failover -n MGMT1
```

The key management node file system resources are `md64` (`/mnt/mgmt`), and `md67` (`/mnt/nfsdata`). By default, the `md64` resource group runs on MGMT0, the `md67` resource group on MGMT1. When MGMT1 is failed over to MGMT0, the `md67` resource is started on MGMT0. Conversely, when MGMT0 is failed over to MGMT1, the `md64` resource is started on MGMT1.

When checking whether MGMT node failover is complete, check that the appropriate `md` resource is started on the partner node.

d. After the failover completes (that is, after the `md67` resource group has started on MGMT0), power off MGMT1:

```
[MGMT0]# pm -0 MGMT1
```

e. Wait until the partner node "`crm_mon`" shows the down node as `OFFLINE`:

```
[MGMT0]# crm_mon -1r | grep -i line
Online:  [ cls12345n000 ]
OFFLINE: [ cls12345n001 ]
```

f. When the node that was shut down shows `OFFLINE`, reboot it:

```
[MGMT0]# pm -1 MGMT1
```

g. Wait until the partner node "`crm_mon`" shows the down node as `ONLINE`:

```
[MGMT0]# crm_mon -1r | grep -i line
Online:  [ cls12345n000  cls12345n001 ]
OFFLINE:
```

h. After MGMT1 reboots, fail back MGMT1:

```
[MGMT0]# cscli failback -n MGMT1
```

i. Verify that failback was successful by checking the location of the `md67` resource group:

```
[MGMT0]# crm_mon -1r
```

j. After the failback completes, fail over MGMT0:

```
[MGMT0]# cscli failover -n MGMT0
```

k. Log in to MGMT1 directly so that your session is not lost when MGMT0 is rebooted.

l. After the failover completes, switch to root user (**`sudo su -`**) and power off MGMT0:

```
[MGMT1]# pm -0 MGMT0
```

m. Wait until the partner node "`crm_mon`" shows the down node as `OFFLINE`:

```
[MGMT1]# crm_mon -1r | grep -i line
Online:  [ cls12345n001 ]
OFFLINE: [ cls12345n000 ]
```

n. When the node that was shut down shows `OFFLINE`, reboot it:

```
[MGMT1]# pm -1 MGMT0
```

o. Wait until the partner node "`crm_mon`" shows the down node as `ONLINE`:

```
[MGMT1]# crm_mon -1r | grep -i line
Online:  [ cls12345n000  cls12345n001 ]
OFFLINE:
```

p. After MGMT0 reboots, fail back MGMT0:

```
[MGMT1]# cscli failback -n MGMT0
```

q. Verify that failback was successful by checking the location of the `md64` resource group.

The system should now be on MGMT0.

r. After the failback completes, log in to MGMT0, change to root, and power on the diskless nodes:

```
[MGMT0]# pm -1 $(nodeattr -s diskless)
```

14. Check that the diskless nodes have rebooted:

```
[MGMT0]# pdsh -g mds,oss "crm_mon -1 | grep Online" | dshbak -c
----------------
cls12345n[002-003]
----------------
Online: [cls12345n003 cls12345n002 ]
----------------
cls12345n[004-005]
----------------
```

```
Online: [cls12345n004 cls12345n005 ]
----------------
cls12345n[006-007]
----------------
Online: [cls12345n006 cls12345n007 ]
```

15. Download and replace the `stripe_cache` script on MGMT01:

After installing the SU, the default stripe cache size will be 4096. However, it needs to be set to 16384. This is done in the this step by downloading a new `stripe_cache` script to replace the default version already stored on the secondary MGMT node (MGMT01). The new script contains the correct options that will be applied automatically during the OST assemble stage of the file system mount. During an OST failover, new values will also be re-calculated and applied.

a. Run the following command to verify the correct settings were applied after mounting the file system (assumes non-failover):

```
[root@cls12345n001 ~]# pdsh -g oss 'lctl get_param ost.OSS.ost_io.threads_max; lctl get_param
obdfilter.*.brw_size; cat /sys/block/md?/md/stripe_cache_size' | dshbak -c
ost.OSS.ost_io.threads_max=768
obdfilter.cls01016-OST0000.brw_size=4
16384
```

The stripe cache size of 16384 is the only parameter change from the default setting of 4096 set by the original `stripe_cache` script.

NOTE: If resulting value is incorrect, proceed with the following steps.

If resulting value is correct, skip to Post-Installation instructions.

b. Download the new `stripe_cache` script from the *Crayport website* to a directory (e.g., `/home/admin/`).

c. Copy the `stripe_cache` script file to the `/tmp` directory on MGMT01:

```
[root@cls12345n000 menu]# scp /home/admin/stripe_cache cls12345n001:/tmp/
stripe_cache 100% 1414 1.4KB/s 00:00
```

d. Log in to MGMT1 via SSH:

```
[root@cls12345n000 menu]# ssh cls12345n001
Last login: Tue Jul 24 19:28:59 UTC 2018 from 172.16.2.3 on ssh
```

e. Copy the new `stripe_cache` script to the same folder as the default version. Type **yes** when asked to overwrite the file, and then press **Enter**.

```
[root@cls12345n001 ~]# cp /tmp/stripe_cache /mnt/nfsdata/images/3.0.0-015.91/
appliance.x86_64/usr/sbin/stripe_cache
cp: overwrite '/mnt/nfsdata/images/3.0.0-.15.91/appliance.x86_64/usr/sbin/
stripe_cache'? yes
```

f. Run the following command, if desired, to verify the correct settings were applied after mounting the file system (assumes non-failover):

```
[root@cls12345n001 ~]# pdsh -g oss 'lctl get_param ost.OSS.ost_io.threads_max; lctl get_param
obdfilter.*.brw_size; cat /sys/block/md?/md/stripe_cache_size' | dshbak -c
ost.OSS.ost_io.threads_max=768
obdfilter.cls01016-OST0000.brw_size=4
16384
```

The stripe cache size of 16384 is the only parameter change from the default setting of 4096 set by the original `stripe_cache` script.

5 Post-Installation Instructions

Prerequisites

System update 3.0.0.SU015 has been completed and the system has been rebooted.

About this task

There are additional system configuration changes that must be made after the system has been restarted. After the cluster restart process is complete, perform the post-installation steps to finish the system update.

Procedure

1. Log in to the primary MGMT node via SSH, then change to root user:

   ```
   remote$ ssh -l admin MGMT0
   [MGMT0]$ sudo su -
   ```

2. Check that all OSS nodes are completely started. Verify that each node pair is listed as `Online`:

   ```
   [MGMT0]# pdsh -g oss "crm_mon -1 | grep Online" | dshbak -c
   ----------------
   cls12345n[004-005]
   ----------------
   Online: [ cls12345n004 cls12345n005 ]
   ----------------
   cls12345n[006-007]
   ----------------
   Online: [ cls12345n007 cls12345n006 ]
   ----------------
   cls12345n[008-009]
   ----------------
   Online: [ cls12345n008 cls12345n009 ]
   ```

3. Log into the MGMT NFS server node (`MGMT1`):

   ```
   [MGMT0]# ssh MGMT1
   ```

4. Change directory to `~admin/300.SU015`:

   ```
   [MGMT1]# cd ~admin/300.SU015
   ```

5. Start a screen session. A log will be kept in `~admin/300.SU015/screenlog.0`:

   ```
   [MGMT1]# screen -L -S update
   ```

   Should the SSH session be interrupted, the update script should continue to run inside the screen session. To reconnect:

a. Log back in to the NFS host node via SSH.

b. Run `screen -ls` to confirm the screen is still active.

c. Run `screen -L -RDS update` to reconnect to the screen session.

6. Run the bundle post-installation:

```
[MGMT1]# ./system-update_csl-3.0.0-015.91.sh post-install
```

The post-install process reports `Done` as the last line of the output.

7. Verify that the system bundle updates have been installed:

```
[MGMT1]# ./system-update_csl-3.0.0-015.91.sh verify
Checking repository validity
Repository definition is present and list of packages in repo matches bundle
Checking if there are any packages that were missed by previous updates...
All available packages are at most recent versions
Checking for unfinished transactions
Following SU packages were installed on MGMT nodes:
cls12345n000: system-update_csl-mgmt-3.0.0-015.91.noarch
cls12345n001: system-update_csl-mgmt-3.0.0-015.91.noarch

Following SU packages were installed onto appliance_mgmt image:
system-update_csl-mgmt-3.0.0-015.91.noarch

Following SU packages were installed onto appliance image:
system-update_csl-node-3.0.0-015.91.noarch

check: SUCCESS
verify: SUCCESS
[MGMT1]#
```

The installation was successful if no errors are reported, the output indicates that the repository definition matches the bundle, all available packages are the most recent, and the package versions are correct.

8. Exit from MGMT1 and return to MGMT0.

```
[MGMT1]# exit
[screen is terminating]

[MGMT1]# exit
[MGMT0]#
```

9. **NOTE:** ISSUE: Adding MMUs to the existing cluster may fail due to incorrect hardware profile.

NOTE: Root or sudo privileges are required for this procedure.

Apply workaround for adding MMUs, if needed:

a. Obtain a copy of `mmu_fix.sh` from Cray.

b. Copy the file to `/home/admin` on both MGMT nodes in the cluster. Run

`scp mmu_fix.sh admin@[MGMT0]:/home/admin scp mmu_fix.sh admin@[MGMT1]:/home/admin`

c. Login to MGMT0. Run

```
ssh -l admin [MGMT0]
```

d. Adjust permissions of `mmu_fix.sh` . Run

```
sudo chmod +x mmu_fix.sh
```

e. Execute the script. Run

```
sude /home/admin/mmu_fix.sh
```

f. Logout from MGMT0 and repeat steps c/d/e on MGMT1.

g. Proceed with adding extra MMUs to the cluster.

Refer to *Verify and Update the Firmware* on page 17 to complete the firmware update process.

6 Verify and Update the Firmware

About this task

Verifies ClusterStor firmware update for 3.0.0.SU015. For each of the firmware checks below, first check the firmware (using the `-c` command line option), and if not correct, update the firmware with the `-u` command line option, then recheck as necessary until firmware is updated.

Refer to the firmware change history in the SU release notes for this version to review which firmware updates are needed.

> **NOTE:** Neo 3.0.0 SU11 and subsequent SUs have depracated and disabled the Xybridge by default. For more information or to request the "ClusterStor: Disabling unused xybridge link on NEO_3.0 systems" article, contact a Cray representative.

Procedure

1. Check and if necessary update the LSI HBA firmware on all nodes:

```
[MGMT0]# export lsitool="/opt/firmware/lsi/release/xrtx_lsifw"

[MGMT0]# pdsh 2>/dev/null -a "${lsitool} -c  && echo 'LSI_FW Needs Update' || \
echo 'LSI_FW is up-to-date'" | egrep FW | sort
cls12345n000: LSI_FW is up-to-date
cls12345n001: LSI_FW is up-to-date
cls12345n002: LSI_FW is up-to-date
cls12345n003: LSI_FW is up-to-date
cls12345n004: LSI_FW is Needs Update
cls12345n005: LSI_FW is up-to-date
```

   Output shows that `cls12345n004` is not updated; run the command with the `-u` command line option to update, then verify with the `-c` command line option. Repeat the sequence until all components are `up-to-date`:

```
[MGMT0]# pdsh -a "${lsitool} -c && (${lsitool} -u && ${lsitool} -a && echo \
'Firmware updated' || echo 'Firmware Not updated' ) || echo \
'Firmware up-to-date'" | dshbak -c
```

2. Check and update the drive firmware:

```
[MGMT0]# pdsh -a "/opt/firmware/drive/release_1/xrtx_drvfw.sh  -c | grep \
Current | cut -f3- -d' ' | sort | uniq -c" | dshbak -c
```

   Ignore:

```
cls12345n002: Can not determine WWN for sda, skipping
cls12345n002: Can not determine WWN for sdb, skipping
cls12345n002: Can not determine WWN for sdc, skipping
cls12345n002: Can not determine WWN for sdd, skipping
```

Be sure the Current and Update values match for each node. If they are different, run the command with the -u command line option to update, then verify with the -c command line option. Repeat the sequence if any components still show a version mismatch.

Update the drive firmware:

```
[MGMT0]# pdsh -g mds=primary,oss=primary "/opt/firmware/drive/release_1/xrtx_drvfw.sh -u"
```

Check the drive firmware:

```
[MGMT0]# pdsh -a "/opt/firmware/drive/release_1/xrtx_drvfw.sh  -c | grep \
Current | cut -f3- -d' ' | sort | uniq -c" | dshbak -c
```

3. Check, and if necessary, update the Mellanox HCA firmware on all nodes; wait until fully completed before continuing with the following step:

   A reboot of all the nodes is required after this update.

   ```
   [MGMT0]# pdsh -a "/opt/firmware/mellanox/release_1/xrtx_mlxfw -c"
   [MGMT0]# pdsh -a "/opt/firmware/mellanox/release_1/xrtx_mlxfw -u"
   cls12345n001: Name: 01:00.0  Current: 2.11.1308 Update: 2.11.1308
   pdsh@cls12345n000: cls12345n001: ssh exited with exit code 11
   cls12345n000: Name: 01:00.0  Current: 2.11.1308 Update: 2.11.1308
   pdsh@cls12345n000: cls12345n000: ssh exited with exit code 11
   ```

   Be sure the Current and Update values match for each node. If they are different, run the command with the -u command line option to update, then verify with the -c command line option. Repeat the sequence if any components still show a version mismatch.

4. Reboot the system if any firmware updates were performed. If firmware was **not** updated on the management nodes, then it is not necessary to reboot the management nodes.

   a. Power off all the diskless nodes:

   ```
   [MGMT0]# pm -0 $(nodeattr -s diskless)
   ```

   b. Verify that the nodes are off:

   ```
   [MGMT0]# pm -q $(nodeattr -s diskless)
   ```

   c. Fail over MGMT1:

   ```
   [MGMT0]# cscli failover -n MGMT1
   ```

   The key management node filesystem resources are md64 (/mnt/mgmt), and md67 (/mnt/nfsdata). By default, the md64 resource group runs on MGMT0, the md67 resource group on MGMT1. When MGMT1 is failed over to MGMT0, the md67 resource is started on MGMT0. Conversely, when MGMT0 is failed over to MGMT1, the md64 resource is started on MGMT1.

   When checking whether MGMT node failover is complete, check that the appropriate md resource is started on the partner node.

   d. After the failover completes (that is, after the md67 resource group has started on MGMT0), power off MGMT1:

   ```
   [MGMT0]# pm -0 MGMT1
   ```

   e. Wait until the partner node "crm_mon" shows the down node as OFFLINE:

```
[MGMT0]# crm_mon -1r | grep -i line
Online:  [ cls12345n000 ]
OFFLINE: [ cls12345n001 ]
```

f. When the node that was shut down shows `OFFLINE`, reboot it:

```
[MGMT0]# pm -1 MGMT1
```

g. Wait until the partner node "`crm_mon`" shows the down node as `ONLINE`:

```
[MGMT0]# crm_mon -1r | grep -i line
Online:  [ cls12345n000  cls12345n001 ]
OFFLINE:
```

h. After MGMT1 reboots, fail back MGMT1:

```
[MGMT0]# cscli failback -n MGMT1
```

i. Verify that failback was successful by checking the location of the md67 resource group:

```
[MGMT0]# crm_mon -1r
```

j. After the failback completes, fail over MGMT0:

```
[MGMT0]# cscli failover -n MGMT0
```

k. Log in to MGMT1 directly (so that your session is not lost when MGMT0 is rebooted).

l. After the failover completes, power off MGMT0:

```
[MGMT1]# pm -0 MGMT0
```

m. Wait until the partner node "`crm_mon`" shows the down node as `OFFLINE`:

```
[MGMT1]# crm_mon -1r | grep -i line
Online:  [ cls12345n001 ]
OFFLINE: [ cls12345n000 ]
```

n. When the node that was shut down shows `OFFLINE`, reboot it:

```
[MGMT1]# pm -1 MGMT0
```

o. Wait until the partner node "`crm_mon`" shows the down node as `ONLINE`:

```
[MGMT1]# crm_mon -1r | grep -i line
Online:  [ cls12345n000  cls12345n001 ]
OFFLINE:
```

p. After MGMT0 reboots, fail back MGMT0:

```
[MGMT1]# cscli failback -n MGMT0
```

q. Verify that failback was successful by checking the location of the md64 resource group.

The system should now be on MGMT0.

r. After the failback completes, log in to MGMT0, change to root, and power on the diskless nodes:

```
[MGMT0]# pm -1 $(nodeattr -s diskless)
```

5. Continue to update the GEM GOBI firmware and the USM firmware, if required. Please contact Cray Support for more information.

7 Restart Lustre

Prerequisites

The system update has been applied and verified.

Procedure

1. Restart the Lustre filesystem:

```
[MGMT0]# cscli mount -f filesystem_name
mount: MGS is starting...
mount: MGS is started!
mount: cls12345 is started on cup[002-003]!
mount: cls12345 is started on cup[004-005]!
mount: All start commands for filesystem cls12345 were sent.
mount: Use "cscli show_nodes" to see mount status.
[MGMT0]#
```

2. Verify that the Lustre filesystem has started on all nodes:

```
[MGMT0]# cscli fs_info
------------------------------------------------------------------------
OST Redundancy style: Traditional Array parity (MDRAID)
Disk I/O Integrity guard (ANSI T10-PI) is not supported by hardware
------------------------------------------------------------------------
Information about "cls12345" file system:
------------------------------------------------------------------------
Node         Role Targets  Failover partner Devices
------------------------------------------------------------------------
cls12345n002 mgs   1 / 1   cls12345n003
cls12345n003 mds   1 / 1   cls12345n002     /dev/md66
cls12345n004 oss   4 / 4   cls12345n005     /dev/md0, /dev/md2, /dev/md4, /dev/md6
cls12345n005 oss   4 / 4   cls12345n004     /dev/md1, /dev/md3, /dev/md5, /dev/md7
```

3. If NXD was disabled at the beginning of the SU installation process, re-enable it now.

```
[MGMT0]# cscli nxd enable
```

Lustre clients may now be mounted.