



## **Installing and Configuring Direct Attached Lustre®**

**S-2541-5101**

---

© 2013 Cray Inc. All Rights Reserved. This document or parts thereof may not be reproduced in any form unless permitted by contract or by written permission of Cray Inc.

---

#### U.S. GOVERNMENT RESTRICTED RIGHTS NOTICE

The Computer Software is delivered as "Commercial Computer Software" as defined in DFARS 48 CFR 252.227-7014.

All Computer Software and Computer Software Documentation acquired by or for the U.S. Government is provided with Restricted Rights. Use, duplication or disclosure by the U.S. Government is subject to the restrictions described in FAR 48 CFR 52.227-14 or DFARS 48 CFR 252.227-7014, as applicable.

Technical Data acquired by or for the U.S. Government, if any, is provided with Limited Rights. Use, duplication or disclosure by the U.S. Government is subject to the restrictions described in FAR 48 CFR 52.227-14 or DFARS 48 CFR 252.227-7013, as applicable.

---

The following are trademarks of Cray Inc. and are registered in the United States and other countries: Cray and design, Sonexion, Urika, and YarcData. The following are trademarks of Cray Inc.: ACE, Apprentice2, Chapel, Cluster Connect, CrayDoc, CrayPat, CrayPort, ECOPhex, LibSci, NodeKARE, Threadstorm. 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.

---

Adobe is a trademark of Adobe Systems, Inc. CentOS is a trademark of the CentOS project. InfiniBand is a trademark of InfiniBand Trade Association. Lustre is a trademark of Xyratex and/or its affiliates. UNIX, the "X device," X Window System, and X/Open are trademarks of The Open Group.

---

#### RECORD OF REVISION

S-2541-5101 Published December 2013 Supports the Cray Linux Environment (CLE) 5.1.UP01 release running on Cray XC30 systems.

---

# Contents

---

	<i>Page</i>
<b>Introduction [1]</b>	<b>5</b>
1.1 Other Related Publications . . . . .	5
<b>Installing DAL on a New System [2]</b>	<b>7</b>
2.1 Before Starting the DAL Installation . . . . .	7
2.2 Load the IMPS Module . . . . .	7
2.3 Build the DAL Image . . . . .	8
2.4 Create the IMPS Config Set . . . . .	8
2.5 Configuring Storage . . . . .	8
2.5.1 Configuring InfiniBand . . . . .	8
2.5.2 Configuring Multipath . . . . .	10
2.6 Provision the DAL Image . . . . .	12
2.7 Configure Lustre . . . . .	12
2.7.1 Install File System Definition Files . . . . .	14
2.7.2 Add Lustre Mount Point to Compute Node Image . . . . .	14
2.8 Create a CPIO Boot Package . . . . .	14
2.9 Boot DAL Service Nodes with DAL CentOS Boot Image . . . . .	14
2.10 Lustre Post Boot Configuration . . . . .	15
2.10.1 Format and Start the Lustre File System . . . . .	15
2.10.2 Add Lustre Entry to /etc/fstab . . . . .	15
2.10.3 Create File System Mount Point . . . . .	15
2.10.4 Mount File System on the Login Node . . . . .	15
2.10.5 Verify Write Access to File System . . . . .	16
2.11 Configuring a Boot Automation File for DAL . . . . .	16
2.11.1 Create Script on Boot Node . . . . .	16
2.11.2 Shutdown the System . . . . .	16
2.11.3 Update Boot Automation File . . . . .	17
2.11.4 Boot Using the Auto Boot File . . . . .	17
2.11.5 Verify Shutdown/Reboot Procedures (Optional) . . . . .	17
2.12 Lustre File System Management . . . . .	18
S-2541-5101	3



# Introduction [1]

---

This guide contains instructions for the installation and configuration of Direct Attached Lustre (DAL) for Cray XC30 systems and is intended for system administrators who are familiar with operating systems derived from UNIX™.

The Direct Attached Lustre file system is optional on Cray XC30 systems. Installation and configuration of DAL differs from that of external Lustre. Because the Lustre community does not support Lustre on SLES-based servers such as the Cray Linux Environment (CLE) operating system distribution, service nodes that support DAL use a CentOS™ operating system running on ramdisk, as opposed to the shared root file system.

Installation and configuration of DAL is facilitated by Cray's Image Management and Provisioning System, a new set of features that changes how software is installed, managed, provisioned, booted and configured. DAL is currently the only Cray product installed using IMPS, however, in future releases, IMPS will support the installation of other Cray products.

An Adobe™ PDF version of this guide is available on the CrayDoc CD and on the CrayPort website at <http://crayport.cray.com>.

## 1.1 Other Related Publications

The following documents contain additional information that may be helpful:

- *Cray Linux Environment (CLE) Software Release Overview Supplement* (S-2497)
- *Installing and Configuring Cray Linux Environment (CLE) Software* (S-2444)
- *IMPS Guide for DAL Installation* (S-0049)
- *Managing System Software for the Cray Linux Environment* (S-2393)
- *Managing Lustre for the Cray Linux Environment (CLE)* (S-0010)



# Installing DAL on a New System [2]

---

This chapter contains the information and instructions that are required to perform an initial installation and configuration of Direct Attached Lustre (DAL) on a Cray XC30 system.

**Note:** Throughout this document, some examples are left-justified to better fit the page. Left justification has no special significance.

## 2.1 Before Starting the DAL Installation

Perform the following tasks before starting the DAL installation.

- **Review release package documentation.** Read the *CLE 5.1.UP01 Release Errata*, *Limitations for CLE 5.1.UP01* 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 Supplement*.

- **Install the Cray Linux Environment (CLE) base operating system and Cray CLE software packages on your system.**

**Important:** Be sure to complete all of Section 5 in *Installing and Configuring Cray Linux Environment (CLE) Software* before performing the instructions in this section.

**Note:** Although Section 6 of *Installing and Configuring Cray Linux Environment (CLE) Software* contains the procedure for configuring Lustre, the following procedure is for configuring Direct Attached Lustre, which provides access to Lustre storage through service nodes on the Cray system.

## 2.2 Load the IMPS Module

You must be logged on to the SMW as `root`.

```
smw:~ # module load imps
```

## 2.3 Build the DAL Image

This creates an image root that will later be provisioned to generate an image for deployment.

```
smw:~ # impscli build image_recipe \  
dal_cle_5.1up01_centos_6.4_x86-64_ari
```

## 2.4 Create the IMPS Config Set

The configuration for DAL nodes resides on the SMW in a config set. The config set must exist and must be configured for the system to operate correctly. Configuration set names for DAL match the name of the partition to be booted, i.e., p0, p1, p2, or p3. For further information on IMPS, see *IMPS Guide for DAL Installation*.

```
smw:~ # impscli create config_set pN with images \  
dal_cle_5.1up01_centos_6.4_x86-64_ari
```

At this point, the IMPS Configurator prompts you for information about your system. A description and guidance, including a reasonable default value, are provided for each query. This process is similar to defining a `CLEinstall.conf`, but done interactively.

The prompts generated by the IMPS Configurator vary from site to site due to system differences and, therefore, are not included in this guide.

## 2.5 Configuring Storage

The IMPS Configurator also prompts for storage details. Key prompts for storage configuration are included in the following sections.

### 2.5.1 Configuring InfiniBand

If InfiniBand storage is connected to the DAL nodes on your system, the following configuration must be performed.



When creating or configuring a system set, you are asked to determine the members of the class `ib-oss`. Any nodes included in this group will have InfiniBand storage configured. If your system has no InfiniBand storage, leave this class empty.

```
#####
```

```
system_config : class : ib-oss : nid
```

```
#####
```

Description:

A list of NID (Node IDentifier) names that provide Lustre via Infiniband (`ib-oss`) services.

Short Description: A comma separated list of `ib-oss` nodes.

Provide a list of all NIDS providing Lustre via Infiniband in this partition; appropriate formats are comma separated NIDS (ex. '1,2,8') and hyphen ranged names (ex. '6-8,12-14'). Mixed single and ranges are supported. If no nodes match the description, leave this entry empty.

Enter string value (press return for default value of ''):

You are prompted for InfiniBand configuration parameters along with a description of the value you are setting and a reasonable default value. These should allow the configuration of most InfiniBand storage subsystems. If your system requires configuration not covered here, please contact your Cray Service personnel and they will assist you with additional configuration options.

```
#####
```

```
system_config : system : class : ib-oss : manipulators : 001 : ensure_equals : files :  
/etc/modprobe.conf.local : options ib_srp srp_sg_tablesize
```

```
#####
```

Description:

This is the maximum number of scatter/gather entries per I/O. A maximum of 255 S/G entries is possible for the InfiniBand SRP driver.

Short Description: This is the maximum number of scatter/gather entries per I/O.

Cray recommends this be left as '255'.

Enter integer value (press return to keep current set value of '255'):

```
#####
system_config : system : class : ib-oss : manipulators : 001 : ensure_equals : files :
/etc/rdma/rdma.conf : IPOIB_LOAD
#####
```

**Description:**

Load the IP over InfiniBand (IPoIB) driver. This driver is required for TCP/IP over InfiniBand support.

**Short Description:** Load the IPoIB driver.

Should be left as 'no' on any OSS node connected to InfiniBand attached storage unless IPoIB connectivity also required.

Enter string value (press return to keep current set value of 'no'):

```
#####
system_config : system : class : ib-oss : manipulators : 002 : ensure_equals : files :
/etc/srp_daemon.conf : a
#####
```

**Description:**

Default SRP target configuration allow rule attribute(s) for a target. Supported attributes are 'max\_cmd\_per\_lun' and 'max\_sect'. 'max\_sect', a decimal number specifying the maximum number of 512-byte sectors to be transferred via a single SCSI command. 'max\_cmd\_per\_lun', a decimal number specifying the maximum number of outstanding commands for a single LUN.

**Short Description:** Recommended allow rule attribute.

Cray recommends this be left as 'max\_sect=8192' to support 1MiB I/Os. Additional examples: 'max\_sect=8192,max\_cmd\_per\_lun=31'.

Enter string value (press return to keep current set value of 'max\_sect=8192'):

## 2.5.2 Configuring Multipath

If your DAL nodes will utilize multipath, the following configuration must be performed. Press return for the default value: False.

```
#####
system_config : system : class : ib-oss : manipulators : 001 : chkconfig : multipathd
#####
```

**Description:**

The multipath service provides a system to provide redundant paths to a device. This service should be enabled IF your system is properly configured to support multipath. May cause the system to be inoperable if multipath is not properly configured.

**Short Description:** Starts multipathd at boot time.

Set to 'true' for systems properly configured for multipath. If you do not know if your system is properly configured for multipath, set to 'false'.

Enter boolean value (press return for default value of 'False'):

You must provide the correct multipath.conf file in your config set. When the multipath.conf file is present, multipath will be started when the DAL nodes are booted.

**Important:** The following requirements exist for multipath configurations:

- The `multipath.conf` file **must** be valid.
- The system **must** be properly configured for multipath. DAL nodes utilizing multipath should be zoned so that these LUNs are not shared with other partitions or non-DAL nodes within the system.

To facilitate creating a valid `multipath.conf` file, a template file is available on the SMW in `/etc/opt/cray/share/config_set/dist.d/multipath.conf.cray` after you perform the configuration step. You must modify this configuration to have the correct values for your system, verify the proper zoning of the system, then copy this file to `/etc/opt/cray/share/config_set/files/class/ib-oss/etc/multipath.conf`. To help you remember this process, the following message will be presented at configuration time.

```
#####
system_config : system : class : ib-oss : manipulators : 001 : copy_files : files :
/etc/opt/cray/share/files/class/ib-oss/etc/multipath.conf
#####
Description:
This option copies site specific multipath.conf from /etc/opt/cray/share/config_set/
files/class/ib-oss/etc/multipath.conf to the value specified by this value.
```

Short Description: DAL multipath.conf destination

Leave this value as `"/etc/"`. After you finish the IMPS configuration step you must copy the `multipath.conf.cray` file into place in the `config_set` and then edit it to reflect the correct values for your system. The example file can be found in `/etc/opt/cray/share/config_set/dist.d/multipath.conf.cray` and you need to copy it to `/etc/opt/cray/share/files/class/ib-oss/etc/multipath.conf`.

Example:

```
smw:/etc/opt/cray/share/p0 # cp dist.d/multipath.conf.cray
smw:/etc/opt/cray/share/p0 # mkdir -p files/class/ib-oss/etc
smw:/etc/opt/cray/share/p0 # cp dist.d/multipath.conf.cray \
files/class/ib-oss/etc/multipath.conf
```

Enter string value (press return to keep current set value of `"/etc/"`):

After the `multipath.conf` file is in place, multipath will be activated at boot time for all nodes in the `ib-oss` class.

## 2.6 Provision the DAL Image

Provisioning prepares the image contents in the proper format for deployment. The provisioned image is stored in the directory `/opt/xt-images/machine-xtrelease-LABEL-partition/nodetype`.

```
smw:~ # impscli provisiondal image \  
dal_cle_5.1up01_centos_6.4_x86-64_ari to  
/opt/xt-images/machine-xtrelease-LABEL-partition/nodetype
```

You are prompted with a series of configuration questions, and then a message similar to the following is displayed after the provisioning completes successfully.

```
INFO - Provisioning of DAL image  
'dal_cle_5.1up01_centos_6.4_x86-64_ari' successful.
```

## 2.7 Configure Lustre

Configuring DAL Lustre nodes is slightly different than on legacy internal Lustre server configurations as the control files are located in the config set on the SMW instead of on the boot node. For more information on configuring Lustre, see *Managing Lustre for the Cray Linux Environment (CLE)*.

For each file system being defined, create and edit a file system definition file, `fs_name.fs_defs`, entering the appropriate node and device information.

```
smw:~ # cd /opt/cray-xt-lustre-utils/default/etc  
smw:~ # cp example.fs_defs fs_name.fs_defs  
smw:~ # vi fs_name.fs_defs
```

Cray recommends updating only the following portion of the `fs_defs` file.

```
# Lustre server hosts to LNET NIDs mapping.
# Multiple lines are additive.
# Use multiple lines with the same nodes if you have several nids for the same
# nodes.
# Use pdsh hostlist expressions.
# i.e. prefix[a,k-l,...] where a,k,l, are integers with k < l, etc
# Each line should have a one-to-one mapping between the nodes and nids.
nid_map: nodes=nid000[27-29,31] nids=[27-29,31]@gni

# Device configuration. Components can be spread across multiple lines.
## node      specifies the primary device host
## dev       specifies the device path
## fo_node   specifies the backup device host
## fo_dev    specifies the backup device path. Only needed if different from
##           the primary device path
## jdev      specifies the external journal device (OST configuration only).
## index     Force a particular OST or MDT index. If this component is specified
##           for one OST or MDT it should be specified for all of them. By
##           default the index is zero based and is assigned based on the order
##           in which devices are defined in this file. i.e. the first 'ost:'
##           has index '0', the second has index '1', the first 'mdt:' has index
##           '0', the second has index '1', etc.
##
##           Note: A combined MGT/MDT target is not supported with multiple MDTs.
##           Note: A separate MGT and MDT can be co-located on a single server.

## MGT
## Management Target
mgt: node=nid00027
    dev=/dev/disk/by-id/scsi-360001ff020021101061ad79111170000

## MDT
## MetaData Target(s)
mdt: node=nid00027
    dev=/dev/disk/by-id/scsi-360001ff020021101061ad79111170100
    index=0
mdt: node=nid00029
    dev=/dev/disk/by-id/scsi-360001ff020021101061ad79111170200
    index=1

## OST
## Object Storage Target(s)
ost: node=nid00028
    dev=/dev/disk/by-id/scsi-360001ff020021101061ad79111170300
    index=0
ost: node=nid00031
    dev=/dev/disk/by-id/scsi-360001ff020021101061ad7a811170400
    index=1
```

**Note:** Be sure to save a copy of the updated `fs_defs` file in another location as the `/opt/cray-xt-lustre-utils/default` link changes with CLE/HSS updates.

## 2.7.1 Install File System Definition Files

Enter the following command for each file system created.

**Note:** You must provide the full path to the `fs_name.fs_defs` file.

```
smw:~ # lustre_control install -I /etc/opt/cray/share/pN/lustre \
fs_name.fs_defs
Performing 'install' from smw at Thu Oct 31 16:16:32 CDT 2013

Parsing file system definitions file: fs_name.fs_defs
Parsed file system definitions file: fs_name.fs_defs
The 'fs_name' file system definitions were successfully installed!
```

## 2.7.2 Add Lustre Mount Point to Compute Node Image

Edit the `fstab` file and add the mount point, as shown, for each file system.

```
smw:~ # vi /opt/xt-images/templates/default-pN/etc/fstab
13@gni:/fs_name /lus/fs_name lustre rw,flock,user_xattr 0 0
```

**Note:** There are additional one-time configuration steps that must be performed on the boot node of the booted system. See [Lustre Post Boot Configuration on page 15](#).

## 2.8 Create a CPIO Boot Package

Invoke the `shell_bootimage_LABEL.sh` script to prepare the CentOS boot package for the system with the specified `LABEL`. This creates a `.cpio` file in the `/bootimagedir` directory. Use the `shell_bootimage` script indicated in the output of `CLEinstall`.

```
smw:~ # /var/opt/cray/install/shell_bootimage_LABEL.sh -c \
-b /bootimagedir/bootimage.cpio \
-d /opt/xt-images/machine-xtrelease-LABEL-partition/nodetype/dal_cle_5.1up01_centos_6.4_x86-64_ari
```

## 2.9 Boot DAL Service Nodes with DAL CentOS Boot Image

Currently all service nodes are running SLES; those intended for DAL must be halted and then rebooted with CentOS.

```
smw:~ # xtnmi --partition pN cnames_of_dal_nodes
```

Where `cnames_of_dal_nodes` is a comma-separated list.

**Important:** Wait sufficient time for the service nodes to halt before proceeding.

```
smw:~ # xtbootsys --partition p1 --reboot -L \
dal_cle_trunk_centos_6.4_x86-64_ari cnames_of_dal_nodes
```

## 2.10 Lustre Post Boot Configuration

For each file system created, perform the following steps from the boot node after the system has been booted with the DAL nodes for the first time **only**.

### 2.10.1 Format and Start the Lustre File System

```
boot:~ # lustre_control reformat -f fs_name
boot:~ # lustre_control start -p -f fs_name
Performing 'start' from boot-pN at Fri Nov 8 13:31:47 CST 2013
Starting filesystem(s):
fs_name
All targets mounted successfully
```

### 2.10.2 Add Lustre Entry to /etc/fstab

For the client mount of Lustre to operate correctly, you must also add the Lustre entry to /etc/fstab in class login. The noauto option should be specified.

```
boot:~ # xtopview -c login
default:// # vi /etc/fstab
13@gni:/fs_name /lus/fs_name lustre rw,flock,user_xattr,noauto 0 0
default:// # exit
```

### 2.10.3 Create File System Mount Point

In an xtopview session, create the /lus/*fs\_name* mount point.

```
boot:~ # xtopview
default:// # mkdir -p /lus/fs_name
default:// # exit
```

### 2.10.4 Mount File System on the Login Node

```
boot:~ # ssh login
login:~ # mount /lus/fs_name
```



**Caution:** Do not continue until the mount command is successful.

## 2.10.5 Verify Write Access to File System

Verify that you have write access to `/lus/fs_name` on the login node by checking the timestamp of `test.txt`.

```
login:~ # touch /lus/fs_name/test.txt
login:~ # ls -la /lus/fs_name
total 12
drwxr-xr-x 3 root root 4096 Nov 8 13:49 .
drwxr-xr-x 3 root root 4096 Nov 8 13:43 ..
drwxr-xr-x 3 root root 4096 Nov 8 13:31 .lustre
-rw-r--r-- 1 root root 0 Nov 8 13:49 test.txt
```

## 2.11 Configuring a Boot Automation File for DAL

Follow the steps in this section to configure and test the boot automation file for DAL. You are asked to shut down your system so that you can test the customized boot automation files.

For more information about boot automation, see the `xtbootsys(8)` man page.

### 2.11.1 Create Script on Boot Node

Create and save the following script on the boot node in `/root/bin/local.dal-opensm`. This script is called by your boot automation file to ensure that DAL nodes running IB are discovering their LUNs upon boot. This script **must** be executable.

```
#!/bin/sh
#
# Local widget to work around opensm startup at boot time on
# dal nodes with Infiniband.
#
pdsh -w cnames_of_dal_nodes "service opensm restart"
```

Where `cnames_of_dal_nodes` is a comma-separated list.

### 2.11.2 Shutdown the System

Use your site-specific procedures to shut down the system. For example, to shutdown using an automation file, type the following:

```
smw:~ # xtbootsys -s last -a auto.xtshutdown
```

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
```



### 2.11.3 Update Boot Automation File

Edit the boot automation file.

```
smw:~ # vi /opt/cray/hss/default/etc/auto.xhostname
```

1. After booting the boot and SDB nodes, add the following line to boot the DAL nodes:

```
lappend actions [list crms_boot_loadfile dal_cle_5.1up01_centos_6.4_x86-64_ari service \
cnames_of_your_DAL_nodes linux]
```

2. Next, add the following line to restart opensm on the DAL nodes that have InfiniBand attached storage:

```
lappend actions { crms_exec_on_bootnode "root" "/root/bin/local.dal-opensm" }
```

3. Add the following line to boot the rest of the service nodes:

```
lappend actions [list crms_boot_loadfile SNL0 service $data(idlist) linux ]
```

This allows time for the DAL nodes with InfiniBand to find all their service devices.

4. Next, add the following line to start Lustre on the DAL nodes:

```
lappend actions { crms_exec_on_bootnode "root" \
"/opt/cray/lustre-utils/default/bin/lustre_control start -f fs_name" }
```

5. Finally, add the following line to mount the DAL clients:

```
lappend actions { crms_exec_on_bootnode "root" \
"/opt/cray/lustre-utils/default/bin/lustre_control mount_clients -f fs_name -w login" }
```

### 2.11.4 Boot Using the Auto Boot File

From this point forward, you can boot your system using the autoboot file you created. This ensures that the entire system boots normally, including DAL service nodes, and to verify that the Lustre file system is installed and working correctly.

1. On the SMW, set the boot image:

```
smw:~ # xtcli part_cfg update pN -i /bootimagedir/image_name.cpio
```

2. Boot the system:

```
smw:~ # xtbootsys --partition pN -a auto.xhostname
```

### 2.11.5 Verify Shutdown/Reboot Procedures (Optional)

Reboot your system and confirm that shutdown and boot procedures operate as expected.

## 2.12 Lustre File System Management

The installation of your Lustre file system is complete. For further information on day-to-day Lustre file system operation and management see *Managing Lustre for the Cray Linux Environment (CLE)*. Additional information can be found in the *Lustre Operations Manual* available at [http://wiki.lustre.org/index.php/Lustre\\_Documentation](http://wiki.lustre.org/index.php/Lustre_Documentation).