Cray DVS Installation and Configuration

Private

S–0005–10

Abstract

Cray DVS Installation and Configuration S–0005–10

This paper provides instructions for installing and configuring the Cray Data Virtualization Service (Cray DVS) on Cray XT systems running UNICOS/lc 2.0. The paper does not describe the design or internal workings of Cray DVS.

Record of Revision

Version	*Description*
1.0	January 2008 Supports limited availability versions of Cray DVS for the UNICOS/lc 2.0 release running on Cray XT systems.

Contents

Introduction [1]

The Cray Data Virtualization Service (Cray DVS) is a distributed network service that provides transparent access to NFS file systems residing on the service I/O (SIO) nodes. Cray DVS provides a service analogous to NFS. The key difference is Cray DVS provides I/O performance and scalability to large numbers of nodes, far beyond the typical number of clients supported by a single NFS server.

The limited availability release of Cray DVS running on the UNICOS/lc 2.0 release provides support for access to NFS file systems. This allows applications running on the compute nodes to read and write data files to the users home directory. Figure 1, page 1 presents a typical Cray DVS use case.

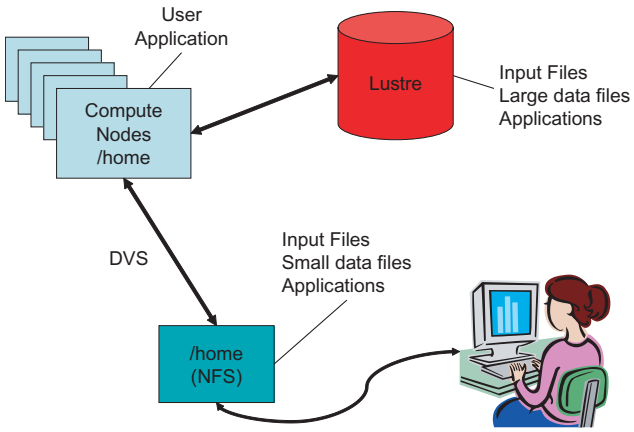

Figure 1. Cray DVS Typical Use Case

For users who are migrating from Catamount to CNL, Cray DVS provides functionality similar to `yod` NFS access on Catamount compute nodes. Normal systems calls such as `open()`, `read()` and `write()` work without modification. Impact on compute node memory resources, as well as operating system jitter, is minimized in the Cray DVS configuration. DVS-specific options to the mount command enable client access to a network file system being projected by DVS server nodes. See the `mount`(8) and `dvs`(5) man pages for more information.

Figure 2, page 2 illustrates the system administrator's view of Cray DVS. Administration of Cray DVS is very similar to configuring and mounting any Linux file system.

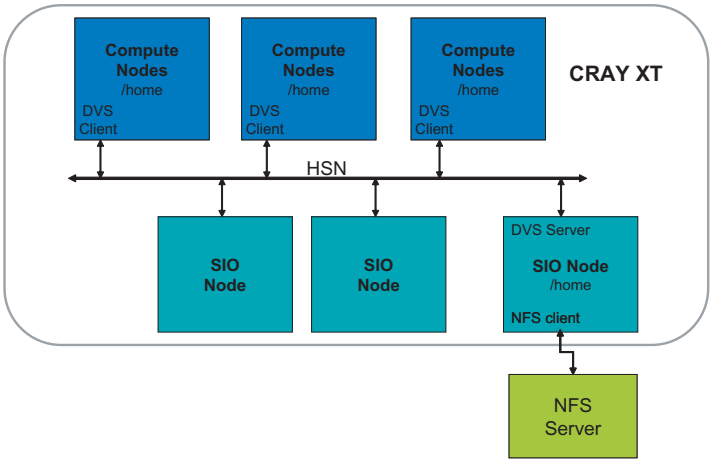

Figure 2. System Administrator's View of Cray DVS

Prerequisites [2]

Before you begin installing and configuring Cray DVS:

- Obtain the Cray DVS RPMs from your Cray representative.
- Your Cray XT system must be running the UNICOS/lc 2.0 release. Verify that the Cray DVS RPMs being installed were generated for the UNICOS/lc 2.0 update level currently running on your system.

Warning: Cray DVS RPMs must be generated to specifically match the UNICOS/lc release level. Customers running the limited availability release of Cray DVS for UNICOS/lc 2.0 will need to request and install updated Cray DVS RPMs each time a new UNICOS/lc update package is installed. Contact your Cray Representative for more information.

- Determine which network file systems will be supported using Cray DVS.
- Determine which SIO node will be configured as the DVS server for each network filesystem. Verify connections to the network file systems on the DVS servers.

Cray DVS Installation [3]

Follow these procedures to install Cray DVS and create a node map file for CNL compute nodes and service nodes. Cray DVS uses the node map file to determine which nodes are participating in DVS communication and where these nodes are located in the mesh.

3.1 Installing the Cray DVS RPMs

These steps assume that you have obtained two RPM files called dvs-ss*.rpm and dvs-cnl*.rpm and copied them to the System Management Workstation (SMW) in a directory called `/tmp/dvs`.

Install the `dvs-ss` RPM on the shared root using the following commands:

```
smw:~> scp -p /tmp/dvs/dvs-ss*.rpm root@boot:/rr/current/software/
smw:~> ssh root@boot
boot001:~# xtopview
default/:/ # rpm -ivh /software/dvs-ss*.rpm
default/:/ # exit
boot001:~# exit
smw:~>
```

Install the `dvs-cnl` RPM to your CNL image using the following commands, where *xthostname-XT_version* is the name of your CNL image:

```
smw:~# cd /opt/xt-images
smw:/opt/xt-images # xtclone xthostname-XT_version xthostname-XT_version-dvs
smw:/opt/xt-images # rpm -ivh --nodeps --root /opt/xt-images/xthostname-XT_vers
/tmp/dvs/dvs-cnl*.rpm
```

3.2 Creating the node-map Files

Once the Cray DVS RPMs have been installed, create the node-map file using the `make-nodemap.sh` script. The `make-nodemap.sh` script creates a node mapping for each node in the Cray XT system, starting at node 0 and moving upward. There are two node-map files created, `node-map.ss` for the DVS SeaStar IPC interface and `node-map.socket` for non-XT systems or systems configured to run DVS over TCP/IP.

Run the **`make-nodemap.sh`** script on the SMW to create the node-map files for the CNL image.

```
smw:~# scp -p
root@boot:/rr/current/opt/dvs/XT_version/usr/sbin/make-nodemap.sh
\
/opt/xt-images/xthostname-XT_version-dvs/etc/dvs
smw:~# cd /opt/xt-images/xthostname-XT_version-dvs/etc/dvs
smw:/opt/xt-images/xthostname-XT_version-dvs/etc/dvs # make-nodemap.sh
smw:/opt/xt-images/xthostname-XT_version-dvs/etc/dvs # ls -l node-map*
lrwxrwxrwx  1 root root    13 Sep  8 05:49 node-map -> ./node-map.ss
-rw-r--r--  1 root root  9012 Sep  8 05:49 node-map.socket
-rw-r--r--  1 root root  5435 Sep  8 05:49 node-map.ss
smw:/opt/xt-images/xthostname-XT_version-dvs/etc/dvs #
```

Install the `node-map.ss` file on the shared root file system.

```
smw:/opt/xt-images/xthostname-XT_version-dvs/etc/dvs
# scp -p node-map.ss \
root@boot:/rr/current/software/
smw:/opt/xt-images/xthostname-XT_version-dvs/etc/dvs # ssh root@boot
boot001:~# xtopview
default/:/ # mkdir /etc/dvs
default/:/ # cp /software/node-map.ss /etc/dvs/
default/:/ # ln -s /etc/dvs/node-map.ss /etc/dvs/node-map
default/:/ # exit
boot001:~# exit
```

Cray DVS Configuration [4]

Follow these steps to configure Cray DVS on your system.

1. For each NFS file system being projected, verify that the DVS server node is running the NFS file system client.
2. Verify that the same directory path exists on the DVS server node that will serve the DVS file system, and that the same directory path exists on the DVS client nodes.
3. Ensure that all DVS server and client nodes have access to an identical (or shared) copy of the `/etc/dvs/node-map` file. This file should include a line for each DVS server and client node.
4. Configure the system to mount the DVS file system on the DVS client nodes by completing the steps in the section below, entitled *Creating fstab Entries and Mount Points.* See the dvs(5) man page for details and examples.
5. Configure a CNL boot image for DVS following the steps in the section below entitled *Creating the Boot Image.*
6. Start the DVS service on all DVS server and client nodes by rebooting the system. The section entitled *Configuring Boot Automation*, describes how to start DVS services automatically.

4.1 Creating `fstab` Entries and Mount Points

After Cray DVS software has been successfully installed on both the service and compute nodes, you can `mount` a network file system on the compute nodes that require access. When a client mounts the file system, all of the information needed is specified on the `mount` command. Follow the steps in this section to configure your Cray XT system to mount a network file system using Cray DVS. See the `dvs`(5) man page for more information regarding Cray DVS mount options.

To allow the compute nodes to mount their DVS partitions, you'll need to add appropriate `fstab` entries. Add a line similar to this example to the `/opt/xt-images/`*`xthostname-XT_version`*`-dvs/etc/fstab` file on the SMW. This example will use DVS to mount `/ufs/home` from node `c0-0c0s1n0` to `/ufs/home` on the client node.

```
smw:~# vi /opt/xt-images/xthostname-XT_version-dvs/etc/fstab
/ufs/home  /ufs/home dvs path=/ufs/home,nodename=c0-0c0s1n0
```

Create mount point directories in the compute image for each DVS mount in the `/etc/fstab` file. For the example shown in the previous section, enter the following command:

```
smw:~ # mkdir -p /opt/xt-images/xthostname-XT_version-dvs/ufs/home
```

Optionally, create any symbolic links that will be used in the compute node images . For example:

```
smw:~ # cd /opt/xt-images/xthostname-XT_version-dvs
smw:~ # ln -s /ufs/home home
```

4.2 Creating the Boot Image

Create the CNL boot image where *parameters* is the path to the parameters list and *BOOTIMAGE* is the either a raw device or the boot image file.

```
smw:~ # xtpackage /opt/xt-images/xthostname-XT_version-dvs
smw:~ # xtbootimg -L /opt/xt-images/xthostname-XT_version-dvs/CNL0.load \
-P parameters -c BOOTIMAGE
```

If *BOOTIMAGE* is a boot image file and not a raw device, update the boot image configuration.

```
smw:~# xtcli boot_cfg update -i BOOTIMAGE
```

The new boot image you have created takes effect when the CNL compute nodes are rebooted.

4.3 Configuring Boot Automation

The `xtbootsys -a` command (see the `xtbootsys`(8) man page) enables you to specify a file to control automated system boot. You can configure this script to start DVS on the SIO nodes responsible for serving DVS file systems. For example, if login nodes are being used to serve the /ufs/home file system, edit the automation file as follows, where *auto.xthostname* is the name of the site specific automation file:

```
smw:~> cd /opt/cray/etc
smw:~> vi auto.xthostname
lappend actions {
{ crms_exec_via_bootnode "login" "root" "/etc/init.d/dvs
start" }
```

Note: DVS should be started on service nodes before booting compute nodes.

After you have configured boot automation, start Cray DVS services by rebooting the Cray XT system.

dvs(5) Man Page [5]

5.1 NAME

dvs — Cray DVS fstab format and options

5.2 SYNOPSIS

```
/etc/fstab
```

5.3 IMPLEMENTATION

UNICOS/lc operating system: supported for Cray XT CNL compute nodes

5.4 DESCRIPTION

The fstab file contains information about which file systems to mount where and with what options. For Cray DVS mounts, the fstab line contains the server's exported mountpoint path in the first field, the local mountpoint path in the second field, and the file system type dvs in the third field. The fourth field contains comma separated DVS-specific mount options described below.

5.5 OPTIONS

path=*/pathname*

Set *pathname* to the mountpoint on the DVS server node. The *pathname* should be an absolute path, and must exist on the DVS server node. This is a required argument on the options field.

`nodename=`*`node`*

Specify the DVS server node name that will provide service to the file system specified by the `path` argument. The path name must exist on the server node specified. Specify the physical ID for the node, for example `c0-0c0s0n0`, which maps to an entry in the `node-map` file where it is translated to a node ordinal. This is a required argument on the options field.

`blksize=`*`n`* Sets the DVS block size to *n* bytes.

`cache` Enables client-side read caching. The client node will perform caching of reads from the DVS server node and provide data to user applications from the page cache if possible, instead of performing a data transfer from the DVS server node.

Note: Cray DVS is not a clustered file system; No coherency is maintained between multiple DVS client nodes reading and writing to the same file. If `cache` is enabled and data consistency is required, applications must take care to synchronize their accesses to the shared file.

`nocache` Disables client-side read caching. This is the default behavior.

`datasync` Enables data synchronization. The DVS server node will wait until data has been written to the underlying media before indicating that the write has completed.

`nodatasync` Disables data synchronization. The DVS server node will return from a write request as soon as the user's data has been written into the page cache on the server node. This is the default behavior.

`retry` Enables the `retry` option, which affects how a DVS client node behaves in the event of a DVS server node going down. If `retry` is specified, any user I/O request is retried until it succeeds, receives an error other than a node down indication, or receives a signal to interrupt the I/O operation. This is the default behavior.

`noretry` Disables the `retry` option. An I/O that failed due to a DVS server node failure will return an `EHOSTDOWN` error to the user application without attempting the operation again.

`clusterfs` Set the `clusterfs` option when the DVS servers are providing access to an underlying file system that is shared or clustered. File I/O to DVS `clusterfs` file systems will go to a single shared file. This is currently the only supported mode for Cray DVS; The `clusterfs` option is set by default.

`maxnodes=`*n* Deferred implementation - this option is not currently supported. If the `clusterfs` option was specified, limit the I/O to a subset of n DVS server nodes out of the list of nodes provided. This allows the administrator to mount a DVS file system that is accessible to a large number of nodes, but have I/O only go to a smaller number nodes out of the possible set. If one of the in-use set of nodes fails, DVS on the client node may choose a replacement node from the larger set.

5.6 EXAMPLES

Here is an example `/etc/fstab` file entry for a DVS client to mount /dvs-shared on the DVS server node as /dvs.

```
/dvs   /dvs   dvs   noauto,path=/dvs-shared,nodename=c0-0c0s1n0
```

5.7 FILES

`/etc/fstab` Static information about file systems

`/etc/dvs/node-map`

Mapping of node ids to node ordinals for DVS

5.8 SEE ALSO

`fstab`(5), `mount`(8), `ummount`(8)