HPC Storage-Lustre Cluster File System Best Particles
Building InfiniBand-Based Lustre Solution

HPC ADVISORY COUNCIL
NETWORK OF EXPERTISE

BEST PRACTICES

The following best practices document is provided as courtesy of the HPC Advisory Council.

To meet the performance, capacity and flexibility requirements an HPC environment requires high-performance storage solutions. The following best practice provides the information for building an InfiniBand-based Luster solution (the described solution is being used by the HPC Advisory Council for applications testing and benchmarking).

The HPC Advisory Council would like to thank Mike Anderson from StreamScale for contributing the InfiniBand-based block storage system that has been built for Lustre. This solution was chosen to be the foundation for the HPC Advisory Council HPC storage system due to its high-performance capabilities and the easy integration with the InfiniBand network that already exists in the Council's HPC Center. If interested, StreamScale solutions can be purchased from multiple vendors, such as Penguin Computing for example

Software and Hardware Components:

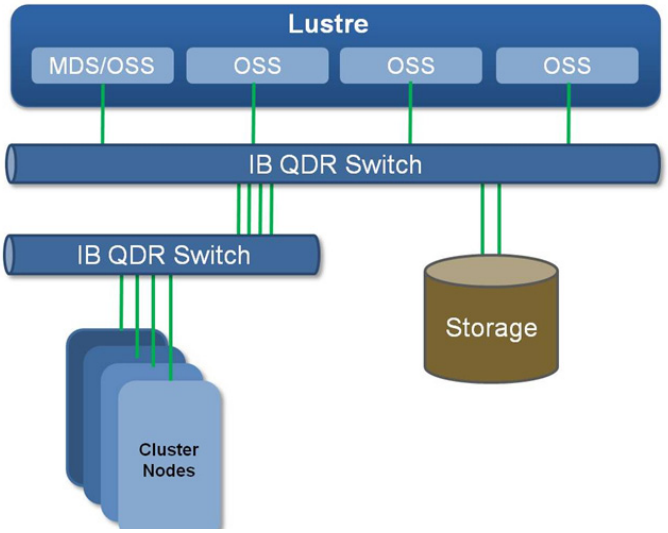
To install Lustre with InfiniBand support, the following are required

Software:

1. Linux kernel patched with Lustre-specific patches (the patched Linux kernel is required only on the Lustre MDS and OSSs)
2. Lustre source code
3. Lustre-specific tools (e2fsck and lfsck) used to repair a backing file system, available in the e2fsprogs package
4. OFED driver

Hardware:

1. Storage Box
   a. Two Intel Core i7 920 CPUs (2.67GHz)
   b. DDR3-1333MHz memory (6GB total)
   c. Seagate Cheetah 15K 450GB SAS Hard Disk
   d. OS: RHEL5.2
   e. Mellanox ConnectX-2 40Gb/s QDR InfiniBand adapter
2. Luster Servers and Lustre Clients
   a. Standard X86 servers
   b. OS: RHEL5.3
   c. Mellanox ConnectX-2 40Gb/s QDR InfiniBand adapter in each server
3. Network Switch
   a. Mellanox M3600 36-Port 40Gb/s QDR InfiniBand switch
4. System Diagram



Configuring InfiniBand based Storage Box

1.Download and install scst-1.0.1.tar.gz from http://scst.sourceforge.net/

The generic SCSI target subsystem for Linux (SCST) allows creation of sophisticated storage devices from any Linux box.

$ tar zxvf scst-1.0.1.tar.gz

$ cd scst-1.0.1

$ make && make install

After install SCST, patch scst.h file with the following content in order to make it interoperate with InfiniBand driver.

```
/*************** Start scst.patch ***************/
        diff -Naur scst/scst.h scst.wk/scst.h
        --- scst/scst.h 2009-08-14 17:09:47.000000000
            -0700
        +++ scst.wk/scst.h      2009-08-14
            17:01:28.000000000 -0700
        @@ -2581,7 +2581,7 @@
         void scst_aen_done(struct scst_aen *aen);

         $if LINUX_VERSION_CODE < KERNEL_VER-
            SION(2, 6, 24)
        -
        +$ifndef RHEL_RELEASE_CODE
         static inline struct page *sg_page(struct scatterl-
            ist *sg)
         {
               return sg->page;
        @@ -2610,6 +2610,7 @@
               sg->length = len;
         }

        +$endif
         $endif /* LINUX_VERSION_CODE < KERNEL_
            VERSION(2, 6, 24) */

         static inline void sg_clear(struct scatterlist *sg)
/****************** End scst.patch ***************/
```

2. Download and Install OFED Driver

Make sure module srpt is loaded by setting in configuration file: ofed.conf

```
$ cd ~/OFED-1.5
$ vi ofed.conf
srptools=y

$ rm RPMS/*
$ ./install.pl -c ofed.conf
```

3. Loading SCST and export hard disk

```
$ modprobe scst
$ modprobe scst_vdisk
$ echo "open vdisk0 /dev/sds BLOCKIO" > /proc/
    scsi_tgt/vdisk/vdisk
$ echo "open vdisk1 /dev/sdu BLOCKIO" > /proc/
    scsi_tgt/vdisk/vdisk
$ echo "add vdisk0 0" >/proc/scsi_tgt/groups/De-
    fault/devices
$ echo "add vdisk1 1" >/proc/scsi_tgt/groups/De-
    fault/devices
```

Installing Lustre from Source Code:

4. Download the latest Lustre kernel and Lustre source code from http://downloads.lustre.org/public/lustre/v2.0/latest/rhel5-x86_64/

5. Install Lustre patched kernel

```
$ rpm -ivh kernel-2.6.18-164.11.1.el5_lus-
    tre.1.10.0.36.x86_64.rpm
$ rpm -ivh kernel-devel-2.6.18-164.11.1.el5_lus-
    tre.1.10.0.36.x86_64.rpm
```

Update the boot loader to boot from the new kernel (/etc/grub.conf)

6. Install the e2fsprogs package

```
$ rpm -ivh e2fsprogs-1.41.6.sun1-0redhat.rhel5.
    x86_64.rpm
```

7. Download the latest OFED and re-build with the new kernel

Configure ipoib address to the InfiniBand interface

8. Install Lustre source code and build Lustre RPMs with InfiniBand support

```
$ ./configure --with-linux=/usr/src/kernels/2.6.18-
    164.11.1.el5_lustre.1.10.0.36-x86_64 --with-o2ib=/
    usr/src/ofa_kernel
$ make rpms
```

RPMs will be generated under /usr/src/redhat/RPMS/x86_64

It is recommended to upgrade all Lustre server and clients to have the same kernel version.

9. Install new Lustre RPMs on all servers

```
$ rpm -ivh lustre-ldiskfs-3.0.9-2.6.18_164.11.1.el5_
    lustre.1.10.0.36_201003050043.x86_64.rpm
$ rpm -ivh lustre-modules-
    1.10.0.36-2.6.18_164.11.1.el5_lus-
    tre.1.10.0.36_201003050042.x86_64.rpm
$ rpm -ivh lustre-1.10.0.36-2.6.18_164.11.1.el5_lus-
    tre.1.10.0.36_201003050042.x86_64.rpm
```

10. Load Lustre network module during every boot on all servers

Add the following lines to the /etc/modprobe.conf file

options lnet networks=o2ib0(ib0)

11. Reboot system

After boot up, the system command lctl list_nids should show: ipadress@o2ib

Configuring Lustre File System:

A Lustre file system consists of four types of subsystems - a Management Server (MGS), a Metadata Target (MDT), Object Storage Targets (OSTs) and clients. It is recommended to run them on a different system.

1. Load Lustre network module during every boot, this needs to be done on all nodes

Adding the following line into /etc/modprobe.conf: options lnet networks=o2ib0(ib0)

2. Configure MDS (Combined MGS/MDT file system)

a. Discover block device (IB storage)

$ modprobe ib_srp srp_sg_tablesize=58

$ ibsrpdm -c > /sys/class/infiniband_srp/srp-mlx4_0-1/add_target

$ ls /dev/sd*

b. Create MGS/MDT on block device:

$ mkfs.lustre –fsname=<fsname> –mgs –mdt <block device name>

$ mkdir <mount point>

c. Mount the combined MGS/MDT file system on the block device

$ mount -t lustre <block device name> <mount point>

3. Configure OSS

$ mkfs.lustre --ost --fsname==<fsname> --reformat --mgsnode=<NID> <block device name>

$ mount -t lustre <block device name> <mount point>

4. Mount Lustre Client

$ mount -t lustre <MGS node>:/<fsname> <mount point>

Verify Lustre file system :

1. Run lfs df -h command on client node

```
lustre-MDT0000_UUID     407.5G    987.3M    383.3G    0% /lustre[MDT:0]
lustre-OST0000_UUID     412.6G     35.3G    356.3G    8% /lustre[OST:0]
lustre-OST0001_UUID     412.6G     31.5G    360.2G    7% /lustre[OST:1]
lustre-OST0002_UUID     412.6G     42.4G    349.2G   10% /lustre[OST:2]
filesystem summary:       3.2T    411.5G      2.7T   12% /lustre
```

HPC ADVISORY COUNCIL
NETWORK OF EXPERTISE

350 Oakmead Pkwy, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.hpcadvisorycouncil.com