Disruptive Storage Workshop
Hands-On Lustre

Mark Miller

http://www.pinedalab.org/disruptive-storage-workshop/

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Schedule

- Setting up our VMs (5 minutes – 15 minutes to percolate)
- Installing Lustre (30 minutes – including 15 minute break)
- Lustre Management - in 2 parts (45 minutes)

Setting up our VMs

Create VM clone for OSS and MDS

- Right-click on “Centos 6.6 Base” VM and select “Clone”
- Name clone “OSS1”
- Make sure you check “Reinitialize the MAC Address”
- Select “Full Clone”
- Once the OSS1 VM is ready, start the VM.

- Right-click on “Centos 6.6 Base” VM and select “Clone”
- Name clone “MDS1”
- Make sure you check “Reinitialize the MAC Address”
- Select “Full Clone”
- Once the MDS1 VM is ready, start the VM.

OSS1 setup

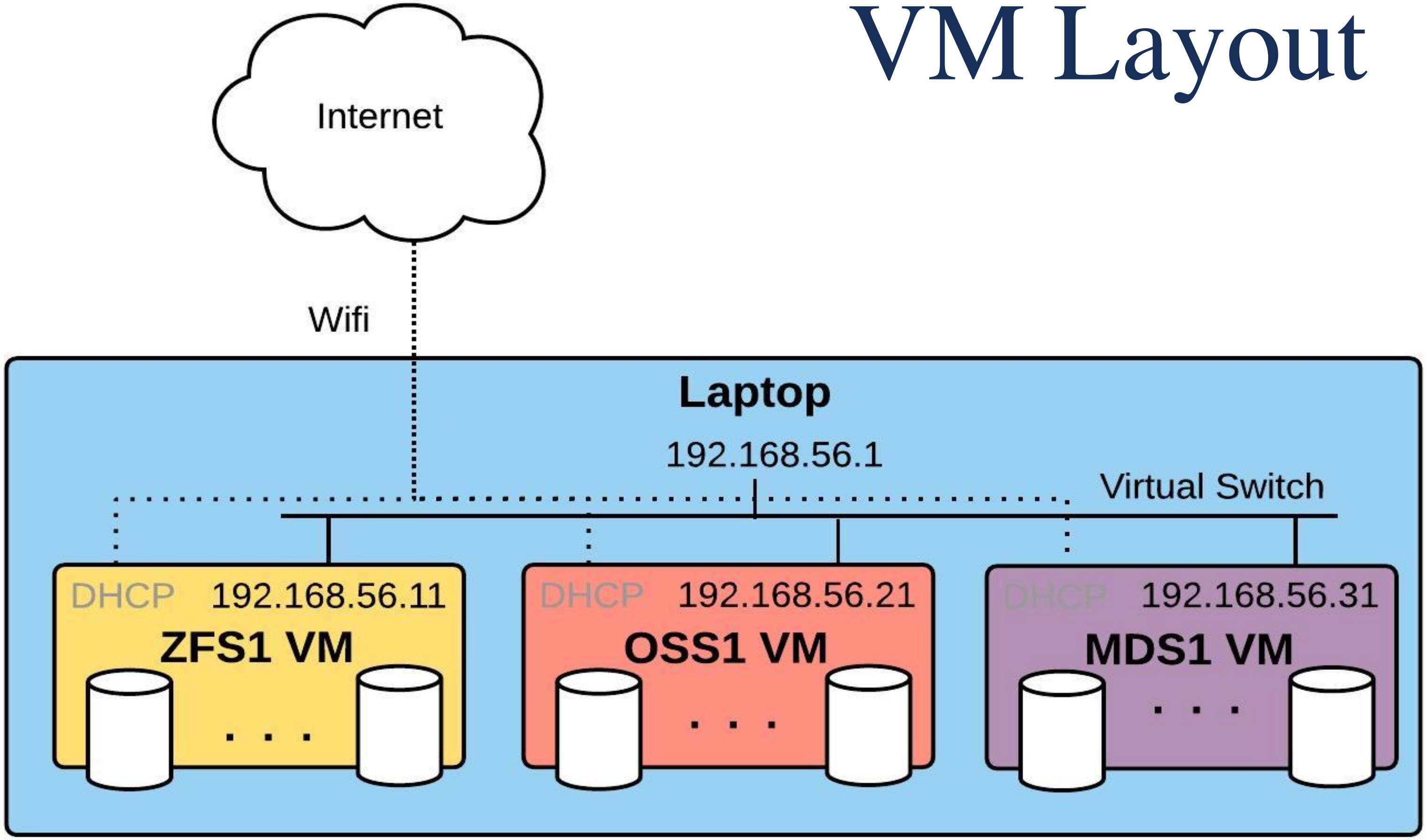
- Login on console: root/password
- cd /software
- Run “./demo-setup oss1”

MDS1 setup

- Login on console: root/password
- cd /software
- Run “./demo-setup mds1”

- The "demo-setup" script will configure the networking on your VM and also setup ZFS. It should take 15 minutes or so to run.

VM Layout



ZFS1 VM will be used at the end for a client
MDS1 VM will house both our MDS and MGS

Let VMs install during next talk….

Installing Lustre

Login via ssh

The IP address for OSS1 is 192.168.56.21
The IP address for MDS1 is 192.168.56.31

Bring up 2 terminals/putty sessions:

```
$ ssh root@192.168.56.21
$ ssh root@192.168.56.31
```

** Please make sure you are not connected to production systems back home.

Now install the Lustre software

As with ZFS, commands for the lab are in `/software/STEPS.lustre`

On both OSS1 and MDS1 the Luster software has been pre-staged in `/software/lustre-server`. To install the Lustre software, on both OSS1 and MDS1:

```
# cd /software/lustre-server
# yum install *.rpm
```

The Lustre packages can also be installed over the Internet from the Intel site with:

```
yum install https://downloads.hpdd.intel.com/public/lustre/lustre-
2.7.0/el6.6/server/RPMS/x86_64/lustre-2.7.0-2.6.32_504.8.1.el6_lustre.x86_64.x86_64.rpm
https://downloads.hpdd.intel.com/public/lustre/lustre-2.7.0/el6.6/server/RPMS/x86_64/lustre-
osd-zfs-2.7.0-2.6.32_504.8.1.el6_lustre.x86_64.x86_64.rpm
https://downloads.hpdd.intel.com/public/lustre/lustre-2.7.0/el6.6/server/RPMS/x86_64/lustre-
dkms-2.7.0-1.el6.noarch.rpm https://downloads.hpdd.intel.com/public/lustre/lustre-
2.7.0/el6.6/server/RPMS/x86_64/lustre-osd-zfs-mount-2.7.0-
2.6.32_504.8.1.el6_lustre.x86_64.x86_64.rpm
```

Lustre packages download location and version information

https://downloads.hpdd.intel.com/public/lustre/lustre-2.7.0/el6.6/server/RPMS/x86_64/

https://wiki.hpdd.intel.com/display/PUB/Lustre+Support+Matrix

A note about the Lustre client RPM

RPMs for the Lustre can be gotten from:

https://downloads.hpdd.intel.com/public/lustre/lustre-2.7.0/el6.6/client/RPMS/x86_64/

… However they do not work on RH 6.7. So the modules had to be built from source:

```
# yum install lustre-client-2.7.0-2.6.32_504.8.1.el6.x86_64.src.rpm
# rpmbuild --rebuild --without servers lustre-client-2.7.0-2.6.32_504.8.1.el6.x86_64.src.rpm
# cp /root/rpmbuild/RPMS/x86_64/lustre-client-2.7.0-2.6.32_573.1.1.el6.x86_64.x86_64.rpm .
# cp /root/rpmbuild/RPMS/x86_64/lustre-client-modules-2.7.0-2.6.32_573.1.1.el6.x86_64.x86_64.rpm .
# yum install *.rpm
```

https://wiki.hpdd.intel.com/display/PUB/Rebuilding+the+Lustre-client+rpms+for+a+new+kernel

http://core.sam.pitt.edu/node/4243

Differences between VM and Real World

- HW/Cabling
- Long Reboot Times
- CPU Affinity for Network Cards
  - Without affinity, "iperf" numbers were 22 Gb/s, With affinity, 32 Gb/s

10 Minute Intermission…

Next Up:

1) Creating the MGS (Management Server)
2) Creating the MDS (Metadata Server)
3) Creating the OSS (Object Stor Server)

Create MGS (Management Server)

Now that the Lustre software is installed, we can create the zpools and Lustre filesystems. First we will create the MGS filesystem. On MDS1, run:

```
zpool create mgs01-pool mirror disk0 disk1

mkfs.lustre --mgs --backfstype=zfs --servicenode=192.168.56.31@tcp1 mgs01-pool/mgs01
```

Create entries in /ets/ldev.conf for this Lustre filesystems:

```
# echo "mds1 - mgs zfs:mgs01-pool/mgs01" > /etc/ldev.conf
```

Check the status of the filesystems:

```
# df -h
# zfs list
```

Create MDS (Metadata Server)

Now that the Lustre software is installed, we can create the zpools and Lustre filesystems. First we will create the MDS filesystem. On MDS1, run:

```
# zpool create mdt01-pool mirror disk2 disk3

# mkfs.lustre --mdt --backfstype=zfs --fsname=lustre01 --index=0 \
    --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.31@tcp1 mdt01-pool/mdt01
```

Create entries in /ets/ldev.conf for these 2 Lustre filesystems:

```
# echo "mds1 - mdt01 zfs:mdt01-pool/mdt01" >> /etc/ldev.conf
```

Check the status of the filesystems:

```
# df -h
# zfs list
```

Configure LNET for our server

The only configuration needed for the Lustre Network (LNET) setup is to configure the network interface and LNET ID that our Lustre Network will use. This is set up in the `/etc/modprobe.d/lustre.conf` file.

```
# echo "options lnet networks=tcp1(eth1)" > /etc/modprobe.d/lustre.conf
```

This is an example of a simple `/etc/modprobe.d/lustre.conf` file, but much more complicated configurations are possible.

Lastly, enable Lustre to startup on boot

```
# chkconfig lustre on
```

Create OSS

Next we will create the OSS zpool and Lustre filesystem. On OSS1, run:

```
# zpool create ost01-pool raidz2 disk0 disk1 disk2 disk3

# mkfs.lustre --ost --backfstype=zfs --fsname=lustre01 --index=1 \
   --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.21@tcp1 ost01-pool/ost01
```

Create entries in /ets/ldev.conf for this Lustre filesystem:

```
# echo "oss1 - oss01 zfs:ost01-pool/ost01" > /etc/ldev.conf
```

Configure LNET for our server:

```
# echo "options lnet networks=tcp1(eth1)" > /etc/modprobe.d/lustre.conf
# chkconfig lustre on
```

Check the status of the filesystems:

```
# df
# zfs list
```

Load Modules and Start Lustre

On MDS1:

```
# modprobe lnet
# lctl dl
# modprobe lustre
# lctl dl
# service lustre start
```

At this point the MDS and MGS are up and running.

```
# lctl dl
# lctl --net tcp1 conn_list
# lctl --net tcp1 list_nids
# netstat -tlpa
```

```
# zfs list
# df -h
```

On OSS1:

```
# modprobe lnet
# lctl dl
# modprobe lustre
# lctl dl
# service lustre start
```

At this point the Lustre service is fully up and running

```
# lctl dl
# lctl --net tcp1 conn_list
# lctl --net tcp1 list_nids
# netstat -tlpa
```

```
# zfs list
# ls /mnt/lustre/local/oss01
```

Install client software on ZFS1

Our Lustre filesystem is now mountable on client systems. The Lustre client needs to be installed, and the Lustre filesystem mounted.

On ZFS1 (192.168.56.11):

```
# cd /software/lustre-client/
# yum install *.rpm

# echo "options lnet networks=tcp1(eth1)" > /etc/modprobe.d/lustre.conf

# mkdir /lustre01
# mount -t lustre 192.168.56.31@tcp1:/lustre01 /lustre01
# df -h
```

Lets do it again!

On MDS1:

```
# mkfs.lustre --mdt --backfstype=zfs --fsname=fritz --index=0 \
    --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.31@tcp1 mdt01-pool/fritz
# echo "mds1 - fritz zfs:mdt01-pool/fritz" >> /etc/ldev.conf
# mkdir /mnt/lustre/local/fritz
# lctl dl
# mount -t lustre mdt01-pool/fritz /mnt/lustre/local/fritz
# lctl dl
```

On OSS1:

```
# zpool create fritz-ost-pool1 raidz2 disk4 disk5 disk6 disk7
# mkfs.lustre --ost --backfstype=zfs --fsname=fritz --index=1 \
   --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.21@tcp1 \
         fritz-ost-pool1/ost1

# echo "oss1 - fritz-ost1 zfs:fritz-ost-pool1/ost1" >> /etc/ldev.conf
# mkdir /mnt/lustre/local/fritz-ost1
# mount -t lustre fritz-ost-pool1/ost1 /mnt/lustre/local/fritz-ost1
```

On ZFS1 (client):

```
# mkdir /fritz-space
# mount -t lustre 192.168.56.31@tcp1:/fritz /fritz-space
# df -h
```

Lustre Management – part 1

How to peek under the hood to see what's going on….

Commands to interact with Lustre

lfs – "User Level" - Lustre Filesystem Management. Only useful where LFS filesystem is mounted (vs MDT/OST)

lctl – "Admin Level" - Lustre Control

The "lfs" commands

- lfs help
- lfs df –h
  - More informative version of traditional df
- lfs check servers
- lfs quota*
  - Allows user based disk quotas
- lfs *stripe* and lfs mv/migrate (in a few slides)

The "lctl" commands

- lctl dl
  - Shows device list

- Display Networking Information
  - lctl --net tcp1 interface_list
  - lctl --net tcp1 peer_list
  - lctl --net tcp1 conn_list

- lct list/get/conf/set_param (operate on /proc)

- lctl lfsck_start / lfsck_stop
  - Status in /proc/fs/lustre/mdd/*/lfsck_layout
  - [root@mds1]# lctl --device fritz-MDT0000 lfsck_start

Files to check for status info

On MDS/MGS:

/var/log/messages
/proc/fs/lustre/osp/*/state
/proc/fs/lustre/mgc/*/state

On OSS:

/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/obdfilter/*/recovery_status

On Client:

/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/osc/*/state
/proc/fs/lustre/mdc/*/state

Shutdown OSS

On client – generate traffic:

```
dd if=/dev/urandom of=/fritz-space/fritz1 bs=1M count=1024 &
```

On OSS – shutdown (power off if needed):

```
init 0
```

On MDS/MGS:

/var/log/messages
/proc/fs/lustre/osp/*/state
/proc/fs/lustre/mgc/*/state

On client:

/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/osc/*/state
NOTE: that the "dd" freezes while OSS is down

Power on OSS

Reboots without errors…?

```
fld: gave up waiting for init of module ptlrpc.
fld: Unknown symbol RQF_FLD_QUERY
...
```

No!!

```
[root@oss1 ~]# echo "modprobe lustre" > /etc/sysconfig/modules/lustre.modules
[root@oss1 ~]# chmod 755 /etc/sysconfig/modules/lustre.modules

[root@mds1 ~]# echo "modprobe lustre" > /etc/sysconfig/modules/lustre.modules
[root@mds1 ~]# chmod 755 /etc/sysconfig/modules/lustre.modules
```

https://jira.hpdd.intel.com/browse/LU-1279
https://jira.hpdd.intel.com/browse/LU-5159

Power on OSS – after module loaded properly

On OSS

- Monitor Console
- more /proc/fs/lustre/obdfilter/*/recovery_status

On MDS/MGS:

```
/var/log/messages
/proc/fs/lustre/osp/*/state
/proc/fs/lustre/mgc/*/state
```

On client:

```
/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/osc/*/state
```

Shutdown MDS

On client – generate traffic:

```
dd if=/dev/urandom of=/fritz-space/fritz1  bs=1M count=1024 &
```

On MDS - shutdown:

```
init 0
```

On OSS:

```
/var/log/messages
/proc/fs/lustre/mgc/*/state
df -h
```

On client:

```
/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/osc/*/state
```

Power on MDS

On MDS/MGS:

```
/var/log/messages
/proc/fs/lustre/osp/*/state
/proc/fs/lustre/mgc/*/state
```

On OSS:

```
/var/log/messages
/proc/fs/lustre/mgc/*/state
```

On client:

```
/var/log/messages
/proc/fs/lustre/mgc/*/state
/proc/fs/lustre/osc/*/state
```

Lustre Management – part 2

How to make changes to our Lustre system.

Add OSS (Part1)

- Clone Centos 6.6 VM to OSS2 VM
- /software/demo-setup oss2 (sets up OS, installs ZFS, installs Lustre)

Add an OST to fritz

To add space, we can add another OST.

On ZFS1 (client)

```
# lfs df -h
```

On OSS1:

```
# zpool create fritz-ost-pool2 disk8 disk9  (why is this a bad idea?)
# mkfs.lustre --ost --backfstype=zfs --fsname=fritz --index=2 \
   --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.21@tcp1 \
         fritz-ost-pool2/ost2

# echo "oss1 - fritz-ost1 zfs:fritz-ost-pool2/ost2" >> /etc/ldev.conf
# mkdir /mnt/lustre/local/fritz-ost2
# mount -t lustre fritz-ost-pool2/ost2 /mnt/lustre/local/fritz-ost2
```

On ZFS1 (client)

```
# lfs df -h
```

Striping data across OSTs

By default, files are written to 1 OST
To stripe files across multiple OSTs:

On ZFS1:

```
# cd /fritz-space
# dd if=/dev/zero of=file1 bs=1M count=100
# lfs getstripe .
# lfs getstripe file1
# lfs setstripe -c 2 .
# dd if=/dev/zero of=file2 bs=1M count=100
# lfs getstripe file2

# lfs mv (for moving a file to a different MDT)
# lfs migrate <file> -o <index>
```

Note df -h on OSS1

Compression

Compression is set at the ZFS level.

On OSS1:

```
[root@oss1 ~]# zfs set compression=on fritz-ost-pool2/ost2

[root@zfs1 fritz-space]# dd if=/dev/urandom of=rand1 bs=1M count=100

[root@oss1 ~]# df -h | grep fritz | grep mnt

[root@zfs1 fritz-space]# dd if=/dev/zero of=zero1 bs=1M count=100

[root@oss1 ~]# df -h | grep fritz | grep mnt
[root@oss1 ~]# zfs set compression=off fritz-ost-pool2/ost2
```

Root Squash

Just like NFS, root-squash can be set up on Lustre. Root sqash is setup on the MDT for a Lustre filesystem

```
# lctl get_param mdt.fritz-MDT0000.nosquash_nids
# lctl get_param mdt.fritz-MDT0000.root_squash

# lctl conf_param  fritz.mdt.nosquash_nids="192.168.56.11@tcp1"
# lctl conf_param  fritz.mdt.root_squash="65534:65534"
```

Lustre Tuning we have done

The only tuning of Lustre that we have done dealt with the "timeout" setting on the clients.

```
[root@compute-091 ~]# lctl get_param timeout
timeout=100
[root@compute-091 ~]# lctl set_param timeout 15
timeout=15
```

Removing an OST

Removing an OST takes several steps on different servers. It requires unmounting the Lustre filesystem from all clients and servers

```
[root@mds1 ~]# lctl dl | grep " osp "
  9 UP osp lustre01-OST0001-osc-MDT0000 lustre01-MDT0000-mdtlov_UUID 5
 16 UP osp fritz-OST0001-osc-MDT0000 fritz-MDT0000-mdtlov_UUID 5
 17 UP osp fritz-OST0002-osc-MDT0000 fritz-MDT0000-mdtlov_UUID 5
[root@mds1 ~]# lctl --device fritz-OST0001-osc-MDT0000 deactivate
[root@mds1 ~]# more /proc/fs/lustre/lov/fritz-MDT0000-mdtlov/target_obd

[root@zfs1 fritz-space]# lfs find --obd 1 /fritz-space/ | lfs_migrate -y
   ? yum install rsync
[root@mds1 ~]# lctl conf_param fritz-OST0001-osc-MDT0000.osc.active=0

[root@zfs1 /]# umount /fritz-space
[root@mds1 ~]# umount /mnt/lustre/local/fritz
[root@oss1 ~]# umount /mnt/lustre/local/fritz-ost1
[root@oss1 ~]# umount /mnt/lustre/local/fritz-ost2

[root@oss1 ~]# tunefs.lustre --ost --backfstype=zfs --writeconf  fritz-ost-pool2/ost2
[root@mds1 ~]# tunefs.lustre --mdt --backfstype=zfs --writeconf  mdt01-pool/fritz

[root@oss1 ~]# mount -t lustre fritz-ost-pool2/ost2 /mnt/lustre/local/fritz-ost2
[root@mds1 ~]# mount -t lustre mdt01-pool/fritz /mnt/lustre/local/fritz
[root@zfs1 /]# mount -t lustre 192.168.56.31@tcp1:/fritz /fritz-space
```

http://wiki.old.lustre.org/manual/LustreManual20_HTML/LustreMaintenance.html#50438199_14978

Adding OSS (part 2)

On MDS1, OSS1, ZFS1, note "lctl dl"
On ZFS1, note "lctl dl", and "df -h"

Login to OSS2 – 192.168.56.22

```
# zpool create fritz-ost-pool3 raidz2 disk1 disk2 disk3 disk4 disk5
# mkfs.lustre --ost --backfstype=zfs --fsname=fritz --index=3 \
   --mgsnode=192.168.56.31@tcp1 --servicenode=192.168.56.22@tcp1 \
         fritz-ost-pool3/ost3

# echo "oss1 - fritz-ost1 zfs:fritz-ost-pool3/ost3" >> /etc/ldev.conf
# mkdir -p /mnt/lustre/local/fritz-ost3
# mount -t lustre fritz-ost-pool3/ost3 /mnt/lustre/local/fritz-ost3
```

oops! Forgot the LNET Entry.

```
# echo "options lnet networks=tcp1(eth1)" > /etc/modprobe.d/lustre.conf
# lustre_rmmod
# mount -t lustre fritz-ost-pool3/ost3 /mnt/lustre/local/fritz-ost3
```

On MDS1, OSS1, ZFS1, note "lctl dl"
On ZFS1, note "lctl dl", and "df -h"

Random Comments

- Commands such as "ls -l" or even "ls" with color enabled can appear slow because those commands need to reach out to the MDS for each file (stat()). Try to encourage "ls --color=none"
- I catch myself spelling words ending in "er" incorrectly, like "clustre"

Thanks for attending! Questions?