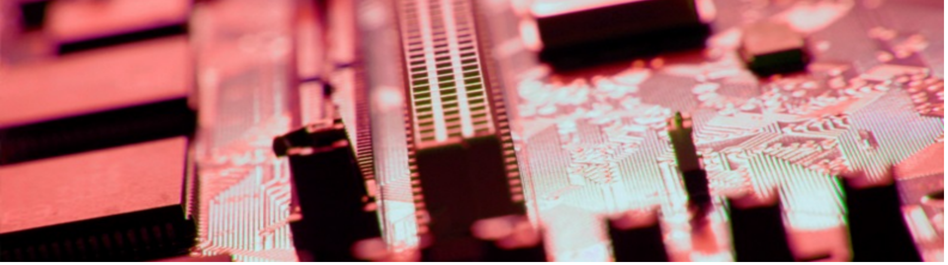


Linux Clusters Institute: Spectrum Scale Hands On Exercise

Georgia Tech, August 15th – 18th 2017

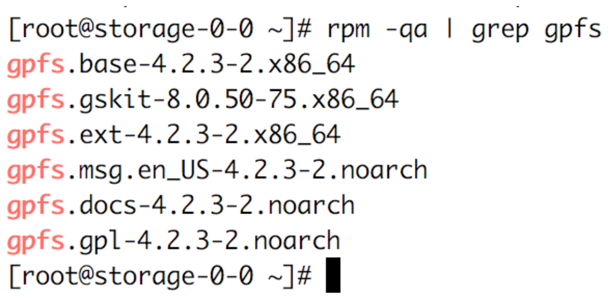
J.D. Maloney | Storage Engineer
National Center for Supercomputing Applications (NCSA)
malone12@illinois.edu

Goal of Hands on Exercise

- Create Spectrum Scale cluster
- Create File System
- Create & Link File sets
- Run policy on sample data
- LUN manipulation
  - Rebalancing/data migration
  - Failure groups
- Explore common commands
  - mmdiag commands
  - mmls* commands

Lay of the Land

- You should have 4 storage servers; 2 for metadata, 2 for data
- Metadata NSD servers have small disks; Data NSD Servers have large disk
- There should be 6 packages installed for Spectrum Scale

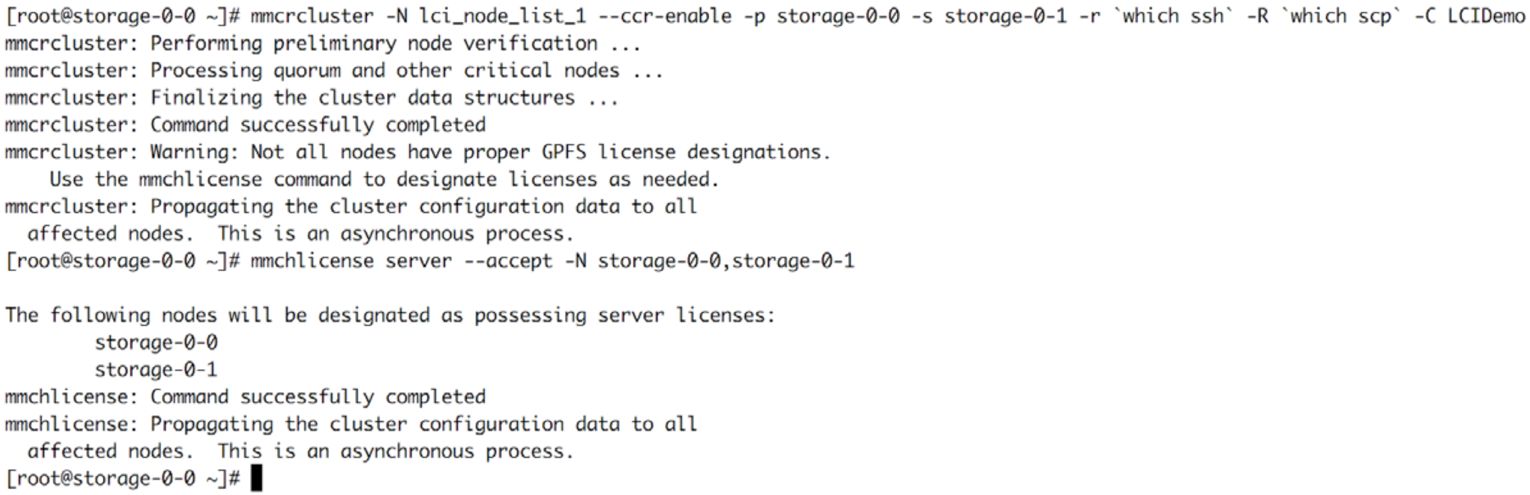


```
[root@storage-0-0 ~]# rpm -qa | grep gpfs
gpfs.base-4.2.3-2.x86_64
gpfs.gskit-8.0.50-75.x86_64
gpfs.ext-4.2.3-2.x86_64
gpfs.msg.en_US-4.2.3-2.noarch
gpfs.docs-4.2.3-2.noarch
gpfs.gpl-4.2.3-2.noarch
[root@storage-0-0 ~]# █
```

- All severs have root ssh keys set between them

Creating Spectrum Scale Cluster

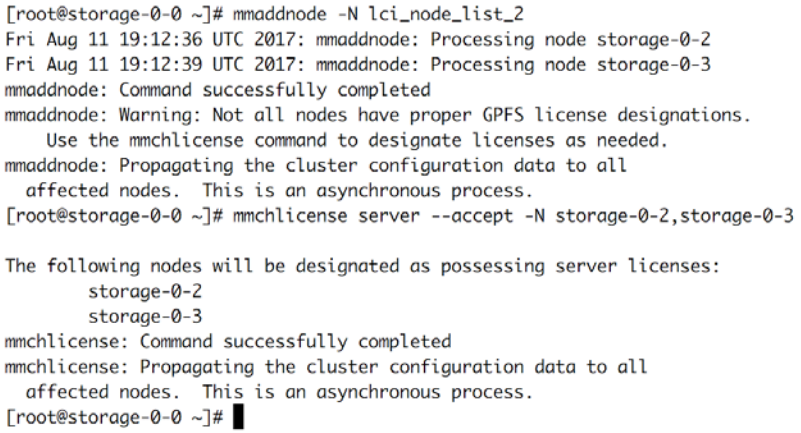
- Decide on key cluster parameters
  - CCR enabled
  - Cluster Name
- Run the create command, use only the two metadata servers to start



```
[root@storage-0-0 ~]# mmcrcluster -N lci_node_list_1 --ccr-enable -p storage-0-0 -s storage-0-1 -r `which ssh` -R `which scp` -C LCIDemo
mmcrcluster: Performing preliminary node verification ...
mmcrcluster: Processing quorum and other critical nodes ...
mmcrcluster: Finalizing the cluster data structures ...
mmcrcluster: Command successfully completed
mmcrcluster: Warning: Not all nodes have proper GPFS license designations.
    Use the mmchlicense command to designate licenses as needed.
mmcrcluster: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# mmchlicense server --accept -N storage-0-0,storage-0-1

The following nodes will be designated as possessing server licenses:
        storage-0-0
        storage-0-1
mmchlicense: Command successfully completed
mmchlicense: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# █
```

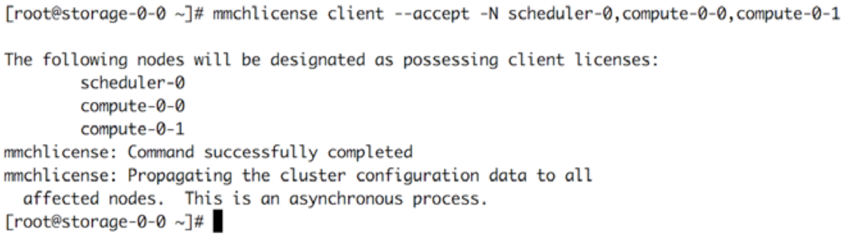
Creating Spectrum Scale Cluster

- Add in your two data NSD servers



```
[root@storage-0-0 ~]# mmaddnode -N lci_node_list_2
Fri Aug 11 19:12:36 UTC 2017: mmaddnode: Processing node storage-0-2
Fri Aug 11 19:12:39 UTC 2017: mmaddnode: Processing node storage-0-3
mmaddnode: Command successfully completed
mmaddnode: Warning: Not all nodes have proper GPFS license designations.
    Use the mmchlicense command to designate licenses as needed.
mmaddnode: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# mmchlicense server --accept -N storage-0-2,storage-0-3

The following nodes will be designated as possessing server licenses:
        storage-0-2
        storage-0-3
mmchlicense: Command successfully completed
mmchlicense: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# █
```

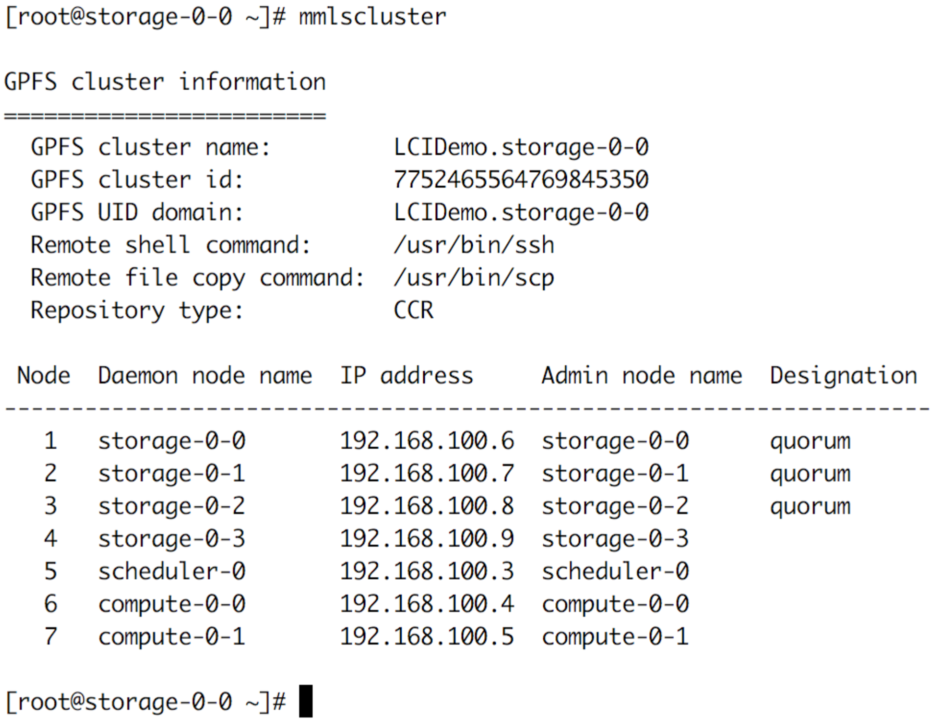
- Now add in your three clients, same command just assign client license



```
[root@storage-0-0 ~]# mmchlicense client --accept -N scheduler-0,compute-0-0,compute-0-1

The following nodes will be designated as possessing client licenses:
        scheduler-0
        compute-0-0
        compute-0-1
mmchlicense: Command successfully completed
mmchlicense: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# █
```

Creating Spectrum Scale Cluster

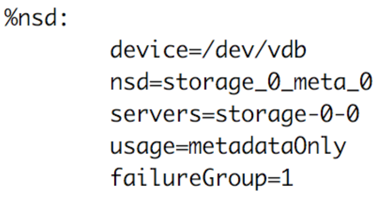
- Verify Cluster



```
[root@storage-0-0 ~]# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         LCIDemo.storage-0-0
  GPFS cluster id:           7752465564769845350
  GPFS UID domain:           LCIDemo.storage-0-0
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address     Admin node name  Designation
--------------------------------------------------------------------
   1   storage-0-0       192.168.100.6  storage-0-0      quorum
   2   storage-0-1       192.168.100.7  storage-0-1      quorum
   3   storage-0-2       192.168.100.8  storage-0-2      quorum
   4   storage-0-3       192.168.100.9  storage-0-3
   5   scheduler-0       192.168.100.3  scheduler-0
   6   compute-0-0       192.168.100.4  compute-0-0
   7   compute-0-1       192.168.100.5  compute-0-1

[root@storage-0-0 ~]# █
```

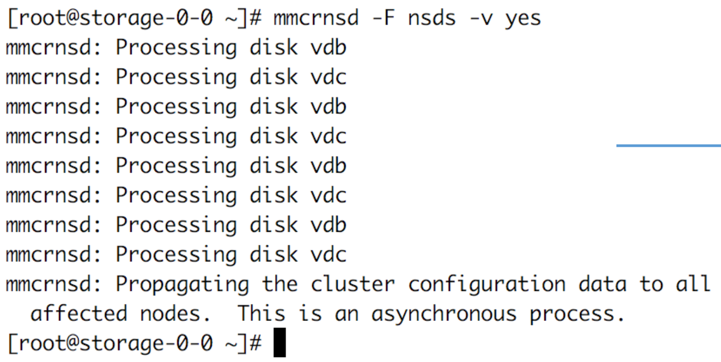
Creating the NSDs

- Create your NSD File
  - Sample stanza below

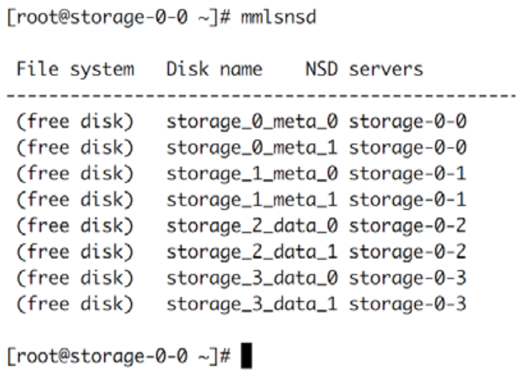


```
%nsd:
        device=/dev/vdb
        nsd=storage_0_meta_0
        servers=storage-0-0
        usage=metadataOnly
        failureGroup=1
```

- Run create command



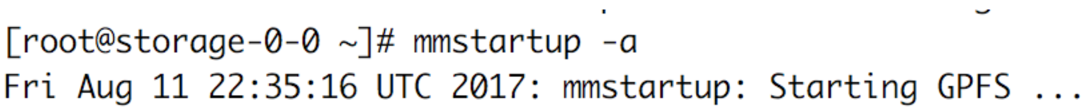
```
[root@storage-0-0 ~]# mmcrnsd -F nsds -v yes
mmcrnsd: Processing disk vdb
mmcrnsd: Processing disk vdc
mmcrnsd: Processing disk vdb
mmcrnsd: Processing disk vdc
mmcrnsd: Processing disk vdb
mmcrnsd: Processing disk vdc
mmcrnsd: Processing disk vdb
mmcrnsd: Processing disk vdc
mmcrnsd: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# █
```



```
[root@storage-0-0 ~]# mmlsnsd

 File system   Disk name        NSD servers
---------------------------------------------------------------------------
 (free disk)   storage_0_meta_0 storage-0-0
 (free disk)   storage_0_meta_1 storage-0-0
 (free disk)   storage_1_meta_0 storage-0-1
 (free disk)   storage_1_meta_1 storage-0-1
 (free disk)   storage_2_data_0 storage-0-2
 (free disk)   storage_2_data_1 storage-0-2
 (free disk)   storage_3_data_0 storage-0-3
 (free disk)   storage_3_data_1 storage-0-3

[root@storage-0-0 ~]# █
```

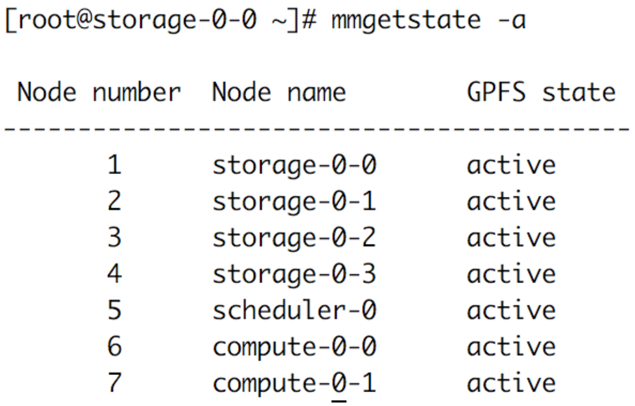
Startup the Cluster

- Startup the cluster on all your nodes



```
[root@storage-0-0 ~]# mmstartup -a
Fri Aug 11 22:35:16 UTC 2017: mmstartup: Starting GPFS ...
```

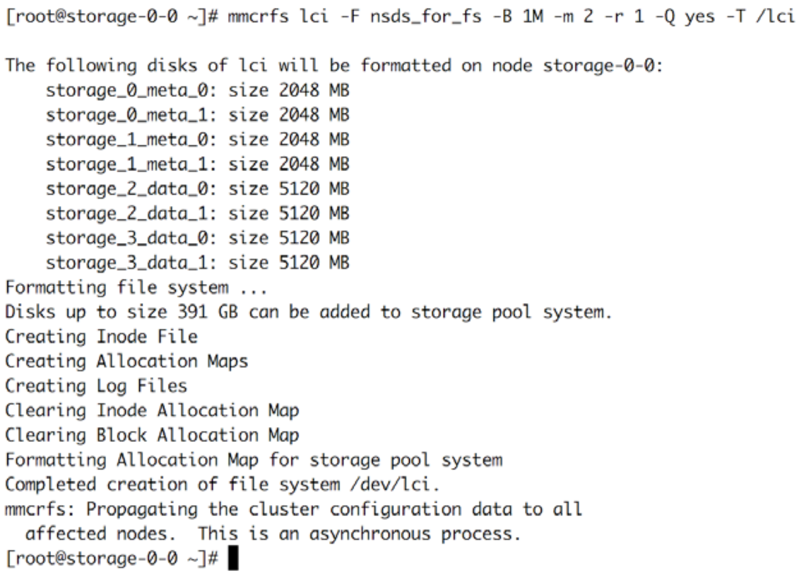
- Wait until they all are in active state



```
[root@storage-0-0 ~]# mmgetstate -a

 Node number  Node name        GPFS state
------------------------------------------
       1      storage-0-0      active
       2      storage-0-1      active
       3      storage-0-2      active
       4      storage-0-3      active
       5      scheduler-0      active
       6      compute-0-0      active
       7      compute-0-1      active
```

Creating the File System

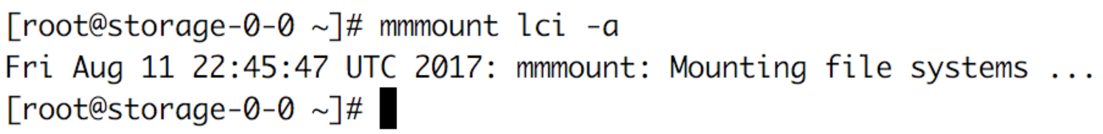
- Create your NSD File
  - Similar to nsd stanza, no device line needed
- Decide on File System Parameters
  - Metadata/Data Replicas: 2/1 respectivey
  - Block size (up to you)
  - Mount Point (up to you)
  - Name (up to you)
- Run create command



```
[root@storage-0-0 ~]# mmcrfs lci -F nsds_for_fs -B 1M -m 2 -r 1 -Q yes -T /lci

The following disks of lci will be formatted on node storage-0-0:
    storage_0_meta_0: size 2048 MB
    storage_0_meta_1: size 2048 MB
    storage_1_meta_0: size 2048 MB
    storage_1_meta_1: size 2048 MB
    storage_2_data_0: size 5120 MB
    storage_2_data_1: size 5120 MB
    storage_3_data_0: size 5120 MB
    storage_3_data_1: size 5120 MB
Formatting file system ...
Disks up to size 391 GB can be added to storage pool system.
Creating Inode File
Creating Allocation Maps
Creating Log Files
Clearing Inode Allocation Map
Clearing Block Allocation Map
Formatting Allocation Map for storage pool system
Completed creation of file system /dev/lci.
mmcrfs: Propagating the cluster configuration data to all
  affected nodes.  This is an asynchronous process.
[root@storage-0-0 ~]# █
```

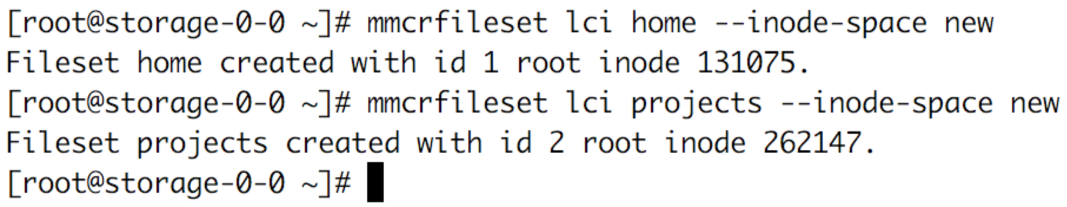
Adding Some Filesets

- Mount your file system on all servers, and your clients



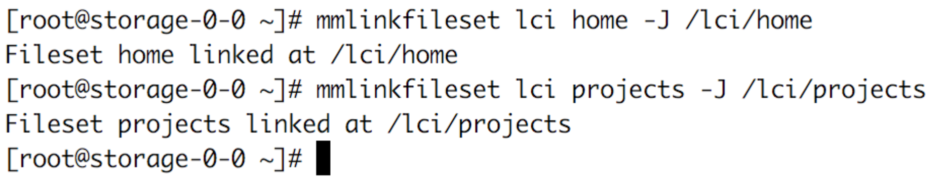
```
[root@storage-0-0 ~]# mmmount lci -a
Fri Aug 11 22:45:47 UTC 2017: mmmount: Mounting file systems ...
[root@storage-0-0 ~]#
```

- Run the mmcrfileset command to two create two filesets



```
[root@storage-0-0 ~]# mmcrfileset lci home --inode-space new
Fileset home created with id 1 root inode 131075.
[root@storage-0-0 ~]# mmcrfileset lci projects --inode-space new
Fileset projects created with id 2 root inode 262147.
[root@storage-0-0 ~]#
```

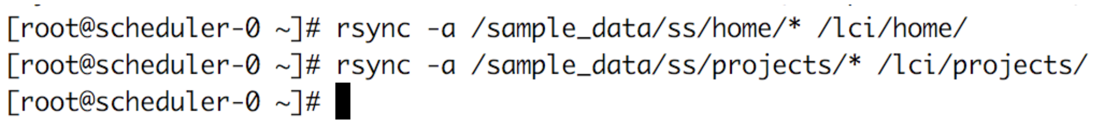
- Link those filesets at the top level of the file system



```
[root@storage-0-0 ~]# mmlinkfileset lci home -J /lci/home
Fileset home linked at /lci/home
[root@storage-0-0 ~]# mmlinkfileset lci projects -J /lci/projects
Fileset projects linked at /lci/projects
[root@storage-0-0 ~]#
```

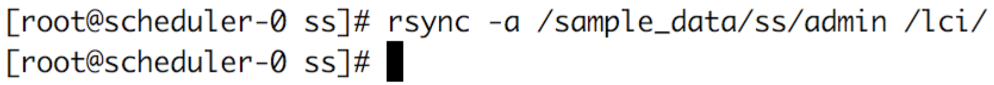
Sample Policy Engine Run

- Pull in sample home & projects data, put in proper dir



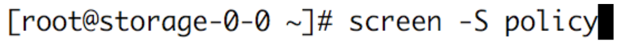
```
[root@scheduler-0 ~]# rsync -a /sample_data/ss/home/* /lci/home/
[root@scheduler-0 ~]# rsync -a /sample_data/ss/projects/* /lci/projects/
[root@scheduler-0 ~]# 
```

- Copy over the sample policy script from lci-sample

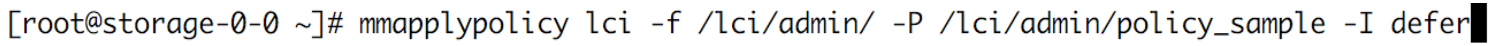


```
[root@scheduler-0 ss]# rsync -a /sample_data/ss/admin /lci/
[root@scheduler-0 ss]# 
```

- Read through policy, discuss it with your team members
- Run policy manually from a screen session



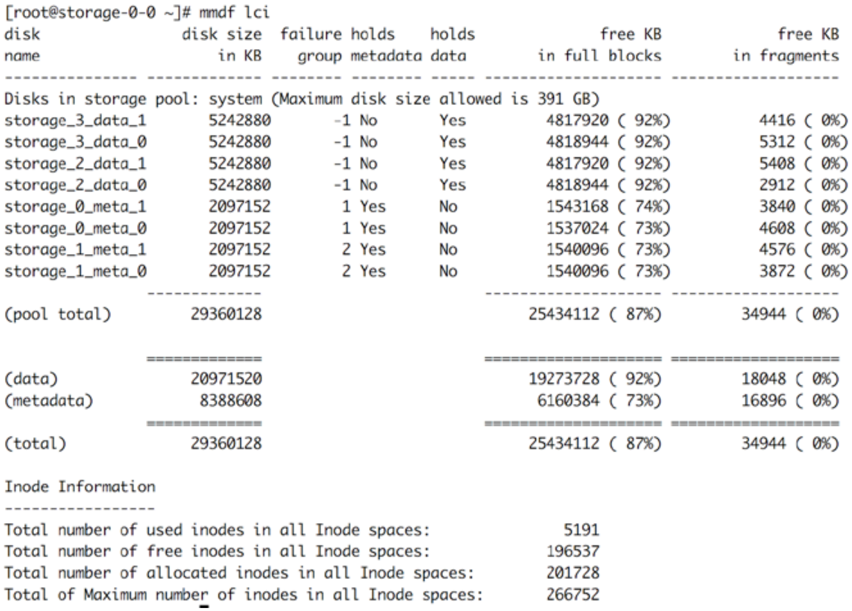
```
[root@storage-0-0 ~]# screen -S policy
```



```
[root@storage-0-0 ~]# mmapplypolicy lci -f /lci/admin/ -P /lci/admin/policy_sample -I defer
```

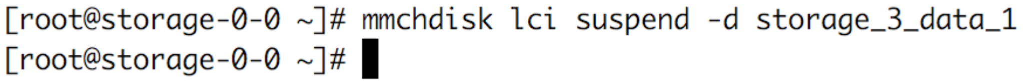
LUN Manipulation

- Take a look at LUN capacities with mmdf (take screenshot)



```
[root@storage-0-0 ~]# mmdf lci
disk                disk size  failure holds    holds              free KB             free KB
name                    in KB    group metadata data        in full blocks        in fragments
--------------- ------------- -------- -------- ----- -------------------- -------------------
Disks in storage pool: system (Maximum disk size allowed is 391 GB)
storage_3_data_1       5242880       -1 No       Yes          4817920 ( 92%)          4416 ( 0%)
storage_3_data_0       5242880       -1 No       Yes          4818944 ( 92%)          5312 ( 0%)
storage_2_data_1       5242880       -1 No       Yes          4817920 ( 92%)          5408 ( 0%)
storage_2_data_0       5242880       -1 No       Yes          4818944 ( 92%)          2912 ( 0%)
storage_0_meta_1       2097152        1 Yes      No           1543168 ( 74%)          3840 ( 0%)
storage_0_meta_0       2097152        1 Yes      No           1537024 ( 73%)          4608 ( 0%)
storage_1_meta_1       2097152        2 Yes      No           1540096 ( 73%)          4576 ( 0%)
storage_1_meta_0       2097152        2 Yes      No           1540096 ( 73%)          3872 ( 0%)
                -------------                         -------------------- -------------------
(pool total)         29360128                              25434112 ( 87%)         34944 ( 0%)

                =============                         ==================== ===================
(data)               20971520                              19273728 ( 92%)         18048 ( 0%)
(metadata)            8388608                               6160384 ( 73%)         16896 ( 0%)
                =============                         ==================== ===================
(total)              29360128                              25434112 ( 87%)         34944 ( 0%)

Inode Information
-----------------
Total number of used inodes in all Inode spaces:               5191
Total number of free inodes in all Inode spaces:             196537
Total number of allocated inodes in all Inode spaces:        201728
Total of Maximum number of inodes in all Inode spaces:       266752
```

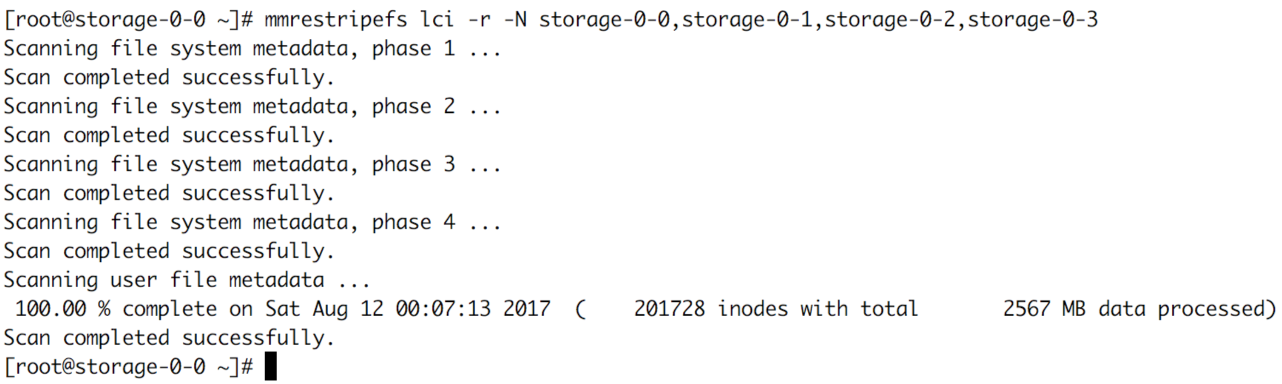
- Suspend one of the data LUNs (doesn't matter which)



```
[root@storage-0-0 ~]# mmchdisk lci suspend -d storage_3_data_1
[root@storage-0-0 ~]#
```

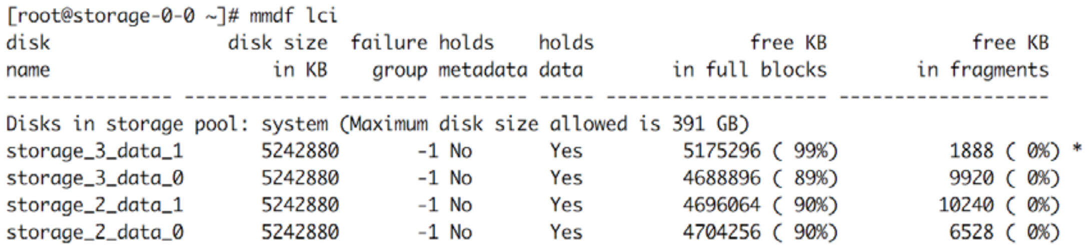
LUN Manipulation

- Run the mmrestripefs commands with the flags you think are appropriate



```
[root@storage-0-0 ~]# mmrestripefs lci -r -N storage-0-0,storage-0-1,storage-0-2,storage-0-3
Scanning file system metadata, phase 1 ...
Scan completed successfully.
Scanning file system metadata, phase 2 ...
Scan completed successfully.
Scanning file system metadata, phase 3 ...
Scan completed successfully.
Scanning file system metadata, phase 4 ...
Scan completed successfully.
Scanning user file metadata ...
 100.00 % complete on Sat Aug 12 00:07:13 2017  (    201728 inodes with total       2567 MB data processed)
Scan completed successfully.
[root@storage-0-0 ~]#
```

- After restripe finishes, run mmdf to verify data left the NSD



```
[root@storage-0-0 ~]# mmdf lci
disk                disk size  failure holds    holds              free KB             free KB
name                    in KB    group metadata data        in full blocks        in fragments
--------------- ------------- -------- -------- ----- -------------------- -------------------
Disks in storage pool: system (Maximum disk size allowed is 391 GB)
storage_3_data_1        5242880       -1 No       Yes          5175296 ( 99%)          1888 ( 0%) *
storage_3_data_0        5242880       -1 No       Yes          4688896 ( 89%)          9920 ( 0%)
storage_2_data_1        5242880       -1 No       Yes          4696064 ( 90%)         10240 ( 0%)
storage_2_data_0        5242880       -1 No       Yes          4704256 ( 90%)          6528 ( 0%)
```

Running Useful Spectrum Scale Commands

- Take a look at the following Spectrum Scale commands, run them, as a team work to understand the output, ask questions if you have them

- mmlsconfig
- mmlscluster
- mmlsmgr
- mmlsnsd
- mmlsdisk
- mmlsfs
- mmlsfileset

mmdiag

Following Flags:

--config

--stats

--network

--waiters

--iohist

Wrap Up

- Further Exploration
  - Other Spectrum Scale Commands you find interesting
- When done
  - Run Through the delete steps
    - mmdelfs
    - mmdelnsd
    - mmdelcluster