GOVERNMENT OF INDIA

INDIRA GANDHI CENTRE FOR ATOMIC RESEARCH

KALPAKKAM - 603102

THIS DOCUMENT IS THE PROPERTY OF INDIRA GANDHI CENTRE FOR ATOMIC RESEARCH
THIS SHALL NOT BE REPRODUCED OR COMMUNICATED WITHOUT OWNER'S PERMISSION

SYSTEM	HIGH PERFORMANCE COMPUTING FACILITY
TITLE	STORAGE SYSTEM CONFIGURATION OF IVY CLUSTER(DDN EXASCALAR)

A	ORIGINAL ISSUE	24/11/2017	
No.	REVISIONS	DATE	APPROVED

COMPUTER DIVISION				**HIGH PERFORMANCE COMPUTING FACILITY**
COMPUTING SYSTEMS SECTION				IGCAR/CD/CSS/HPC/12
	NAMES	SIGN	DATE	
PREPARED	**SUBRAHMANYAM DNVR**	D.N.V.R. Subrahmanyam	24/11/2017	
CHECKED	**SUJA RAMACHANDRAN**	Suja	24/11/2017	
REVIEWED	**M.L.JAYALAL**		24/11/2017	REV A
APPROVED	**R.JEHADEESAN**		24/11/17	Distribution: Head CSS, Head CD, Director,EIG

SUMMARY

The high performance computing facility at Computer Division has multiple cluster computers for meeting the computational requirements of scientists and engineers of IGCAR. Ivy Cluster is recent among them and this document gives the configuration details of Ivy Cluster storage system. The hardware and software configuration details of the storage system are covered in the document. The details about the Lustre file system are also covered.

Storage Configuration of Ivy Cluster

1 Introduction

In a computing cluster, usually there are hundreds of compute nodes present and the HPC applications running on these nodes write to the file system in parallel. Therefore, the storage of a cluster has to be able to handle fairly large number of file IO operations in parallel. Apart from the large number of IO operations (~10,000 IOPs) each operation itself is a large chunk of data. Therefore, the throughput (i.e. the number of bytes transferred in a second) of the storage is another important parameter for assessing the performance of computing cluster storage. Yet another requirement is that the storage should provide an indexed access to the files rather than conventional tree like nested directories based approach to store the files. This helps to reduce delays in traversing the hierarchy of directories while locating a file. Due to these requirements the hardware & software configuration and the file system on the storage have to be special for cluster computers.

Ivy Cluster is one of the large clusters at the HPC facility with a Peak (theoretical) performance of 207 TeraFLOPS and maximal sustained performance of around 180 TeraFLOPS with industry-standard HPL benchmark. It has 400 compute nodes, two GPU nodes and two Phi nodes, all of which have two shared Lustre file systems *'/home'* & *'/apps'* of size 256TB each and with a support of up to 16GBps throughput. The following document describes the hardware and the software configuration of the storage of the Ivy Cluster.

2 Introduction to Lustre

The Ivy Cluster uses Lustre file system(Version: 2.5.29.ddnpf4). Therefore, an overview about the Lustre file system is given in the following description:

Lustre is a client-server based parallel network file system. It separates metadata(inodes) from data storage(file contents). The Lustre servers manage the presentation of storage to the networked clients. Clients aggregate name space and object data to present a coherent POSIX file system to the applications. Any Lustre file system comprises of three different types of services:

1. **Management service** provides the configuration information regarding one or more Lustre file systems. A server running this service is called **Management Server(MGS)** and it uses **Management Target(MGT)** to store this information.
2. **Meta Data Service** serves read/write of the file system metadata like inodes & file system index. A server running this service is called **MetaData Server(MDS)** and it uses **Meta Data Target(MDT)** to store the metadata.
3. **Object Storage Service** serves the read/write of file contents as binary objects, distributed across multiple targets in parallel. A server running this service is called **Object Storage Server(OSS)** and it uses **Object Storage Target(OST)** to store the data.

In case of a minimal Lustre configuration, at least two servers are required where MGS and MDS can be clubbed and run on a single server while OSS can run on a dedicated server. Lustre also supports High Availability(HA) configuration. For HA at least four servers are required. Two Metadata Servers, running MGS and MDS in failover configuration with shared storage for MGT and MDT and two Object storage servers with multiple OSTs configured in failover configuration. The overall architecture of the Lustre file system is depicted in Figure 1: Overview of the Lustre file system.

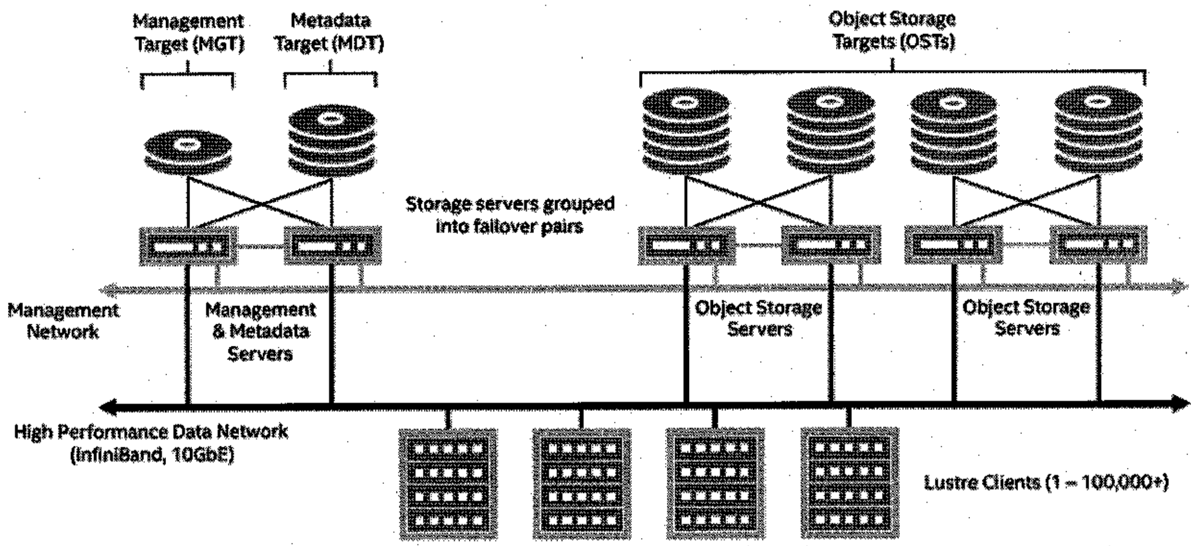
Management Target (MGT)
Metadata Target (MDT)
Object Storage Targets (OSTs)
Storage servers grouped into failover pairs
Management Network
Management & Metadata Servers
Object Storage Servers
Object Storage Servers
High Performance Data Network (InfiniBand, 10GbE)
Lustre Clients (1 – 100,000+)

Figure 1: Overview of the Lustre file system

3 Hardware Configuration

The storage of Ivy cluster is 512 TB (usable) in size and delivers 16GBps of throughput. It has various entities like disks, enclosures, controllers, OSSs and MDSs. Disks are used to store the data and metadata; enclosures house these disks and are connected to the controllers. Controllers provide a block-level access of the underlying storage to the storage servers(OSS or MDS) in the form of Virtual Disks. A virtual disk is a combination of disks in the form of a RAID array. Each of the virtual disk forms a single target. It is called OST if it is used for storing data(like the content of files stored on Lustre file system) and it is called MDT if it is used for storing metadata(like inodes, file names, permissions, references to data on OST etc of the Lustre files). The OSS and MDS servers run the Lustre software components which provide a mechanism to export the Lustre file system over the network. This exported Lustre file system is then mounted on all the compute nodes with the help of Lustre client modules like the LNet driver, Object Storage Client(OSC) and Meta Data Client(MDC) running on each of the compute nodes. User applications running on the compute nodes have access to the files on the Lustre file system similar to that of the files available on the local storage.

Disks are mounted in the enclosure through SAS bus ports. Connectivity between the enclosure and the controller is through dedicated HBA(Host Bus Adapter) link. Connectivity between the OST controllers and the OSS servers is through InfiniBand (IB) network, while connectivity between the MDT Controller and the MDS is through dedicated redundant fiber optic links. The Connectivity between the storage servers OSSs and MDSs and the compute nodes is also through IB network.

3.1 Disks

3.1.1 For data storage:

There are 240 SAS disk drives on the whole, each of 3TB size running @ 7.2K RPM for storing data. These disks are connected to the 2x DDN SFA7700X Storage Controllers which are in-turn connected to the OSS servers through IB network.

3.1.2 For meta-data storage:

There are 12 SAS disk drives each of 600GB size running @ 15K RPM for storing metadata. These disks are connected to the DDN EF3015 Controller which is in-turn connected to MDS servers through redundant fiber optic interface links.

3.2 Enclosures

3.2.1 For data storage:

The number of disks required for data storage is large i.e. 240. Therefore these disks are mounted using enclosures. And there are four enclosures on the whole for OSTs. Each enclosure is a 3U chassis mounted in a 19" rack and is capable of containing 60 disks. Two of the enclosures also contain 1xSFA7700X controller (primary/secondary pair) per enclosure apart from the disks.

3.2.2 For meta-data storage:

For MDT there is a single enclosure which houses all the 12 physical disks along with the DDN EF3015 controller in a 2U chassis mounted in a 19" rack.

3.3 Controllers

3.3.1 Controller for data storage (DDN SFA7700X):

The SFA controller is the interface between the data disks and the object storage servers. It takes care of the RAID configuration implemented in the hardware level. There are two SFA controllers(SFA1 & SFA2) on the whole for OSTs. Each controller is again having Primary and Secondary controller internal to itself for the sake of redundancy.

Access to the SFA controller is possible through SSH session. The IP address of the SFA controller is 30.0.6.19 for SFA1 & 30.0.6.23 for SFA2. The IP addresses of corresponding secondary controllers are 30.0.6.21 & 30.0.6.25. The detailed IP address configuration is furnished in Table 1 below.

Controller	Primary/Secondary	Name	eth0 IP Address	IPMI/BMC
SFA 7700X – 1 (SFA1)	Primary	IGCAR_SFA1 RAID[0]	30.0.6.19	30.0.6.20
	Secondary	IGCAR_SFA1 RAID[1]	30.0.6.21	30.0.6.22

SFA 7700X – 2 (SFA2)	Primary	IGCAR_SFA2 RAID[0]	30.0.6.23	30.0.6.24
	Secondary	IGCAR_SFA2 RAID[1]	30.0.6.25	30.0.6.26

Table 1: IP addresses of SFA controllers

There are two Inter Controller Links (ICLs) between the primary and secondary controller which are used for exchanging heartbeat messages. One is a high speed PCIe link inside to enclosure itself. It is used also to synchronize the cache contents of the primary with secondary controller. Other is an ethernet link. These links are required for configuring High Availability with in a controller. Similar setup is followed for SFA2 as well.

3.3.2 Controller for meta-data storage (DDN EF3015):

Similarly to that of data, EF3015 is the controller used for meta data storage and it connects the metadata disks to the MDS servers. This controller also has a built-in secondary controller for the sake of redundancy. Refer the Table 2 below for IP address details of this controller.

Controller	Primary/Secondary	Name	eth0 IP Address	IPMI/BMC
EF 3015	Primary	IGCAR-MDT_CTRL	30.0.6.27	-
	Secondary	IGCAR-MDT_CTRL	30.0.6.28	-

Table 2: IP addresses of EF3015 Controller

3.4 RAID Configurations

3.4.1 RAID configuration for data storage:

Aggregation of physical disks into storage pools is carried out with appropriate RAID scheme. RAID level 6 is the configuration used in Ivy cluster with 8 disks for data and 2 disks for parity in each RAID array. Multiple RAID Arrays/groups of this configuration are created in the storage. Each group is called as a pool or virtual disk. Since two disks are for parity the total capacity of each pool or a virtual disk turns out to be eight(i.e eight number of data disks) times of the size of single disk(i.e 3TB). Therefore, each pool is of 24TB(22TB usable) capacity. This is the size of the virtual disk that the controller exports to the object storage servers. 12 such virtual disks are exported from each SFA7700X controller to a pair of OSSs. Out of these 12, 6 virtual disks are used by *'/home'* file system and rest of the 6 are used by *'/apps'* file system(*'/home'* & *'/apps'* Lustre file systems are discussed later in this document). Same is done on the other controller as well. Therefore, on the whole there are 24 OSTs.

3.4.2 RAID configuration for meta-data storage:

In order to achieve redundancy for metadata, RAID10(i.e. mirroring with stripping) is used. Therefore, the effective storage capacity becomes six(i.e six number of data disks excluding mirror disks) times of the space size of each disk(i.e 600GB) which becomes 3.6TB. This is configured as a single virtual disk(named: vd01) on the MDT controller(DDN EF3015). But this virtual disk is again partitioned in to 2 physical volumes on the same controller. One is **mdt-apps** with the size of 1751 GB and the other is **mdt-home** with the size of 1845 GB. These two volumes are then exported to the MDSs through dual redundant 8Gbps FC links and are configured as multipath devices. The multipath configuration is described in the following section.

3.5 Virtual disk visibility and multipath configuration on storage servers

3.5.1 Multipath configuration for data storage:

There are 24 virtual disks created in the SFA controller by combining 10 physical disks in 8D+2P RAID6 configuration, as explained earlier. These are exported to the OSSs through the IB network. A process named *'srp_daemon'* is responsible to detect the virtual disks on the OSSs. Once this process is run, the devices can be seen by issuing the command *'lsscsi'* on the storage server. They appear as *'/dev/sdb'*, *'/dev/sdd'* etc. Since there are two IB interfaces to the controller and we want to use these as redundant paths to the storage (for HA), *'srp_daemon'* is configured to utilize both the paths to the same storage controller thereby to the same virtual disks. Therefore, each virtual disk exported on the controller is visible to the storage server as two separate disks(one from each path). These two devices are then configured to form a single multipath device and a user friendly name is assigned to the same. The detailed configuration is mentioned in /etc/multipath.conf file in OSSs. The multipath devices are named sequentially from *'/dev/mapper/ost_home_0' to '/dev/mapper/ost_home_11'* for home file system and similarly *'/dev/mapper/ost_apps_0'* to *'/dev/mapper/ost_apps_11'* for apps file system.

The object storage servers mount these 24 multipath devices(OSTs). There are four OSSs on the whole in Ivy cluster. Each of these OSS mounts 3 home OSTs and 3 apps OSTs. For example, OSS1 mounts */dev/mapper/ost_home_0*, */dev/mapper/ost_home_1* and */dev/mapper/ost_home_2* plus */dev/mapper/ost_apps_0*, */dev/mapper/ost_apps_1* and */dev/mapper/ost_apps_2*. The details about the virtual disks exported from SFA-1 and SFA-2 controllers are shown in Table 3. The details given in the table are for SFA 7700x-1 controller.

Virtual Disks exported from SFA 7700X controller (show vd)	Virtual Disks as they appear to OSSs (lsscsi)	Configured multipath device/alias (multipath –ll)	Mounted on OSS(mounted as)	For file-system
vd-0_0	/dev/sdc /dev/sdo	/dev/dm-8 (ost_home_0)	oss1 /Lustre/home/ost_0	/home
vd-1_1	/dev/sdd /dev/sdp	/dev/dm-12 (ost_home_1)	oss1 /Lustre/home/ost_1	/home
vd-2_2	/dev/sde /dev/sdq	/dev/dm-7 (ost_home_2)	oss1 /Lustre/home/ost_2	/home
vd-3_3	/dev/sdf /dev/sdr	/dev/dm-15 (ost_apps_0)	oss1 /Lustre/apps/ost_0	/apps
vd-4_4	/dev/sdg /dev/sds	/dev/dm-4 (ost_apps_1)	oss1 /Lustre/apps/ost_1	/apps
vd-5_5	/dev/sdh /dev/sdt	/dev/dm-10 (ost_apps_2)	oss1 /Lustre/apps/ost_2	/apps
vd-6_6	/dev/sdi /dev/sdu	/dev/dm-11 (ost_home_3)	oss2 /Lustre/home/ost_3	/home

vd-7_7	/dev/sdj /dev/sdv	/dev/dm-14 (ost_home_4)	oss2 /Lustre/home/ost_4	/home
vd-8_8	/dev/sdk /dev/sdw	/dev/dm-6 (ost_home_5)	oss2 /Lustre/home/ost_5	/home
vd-9_9	/dev/sdl /dev/sdx	/dev/dm-13 (ost_apps_3)	oss2 /Lustre/apps/ost_3	/apps
vd-10_10	/dev/sdm /dev/sdy	/dev/dm-9 (ost_apps_4)	oss2 /Lustre/apps/ost_4	/apps
vd-11_11	/dev/sdn /dev/sdz	/dev/dm-5 (ost_apps_5)	oss2 /Lustre/apps/ost_5	/apps

Table 3: Virtual Disks exported from SFA-1 and SFA-2 controllers

The details for SFA7700X-2 can be derived from Table 3, in a similar fashion.

The following Table 4 shows the summary of which OSTs are mounted on each storage server:

Server	OSTs for /home fs	OSTs for /apps fs
OSS1	ost_home_0	ost_apps_0
	ost_home_1	ost_apps_1
	ost_home_2	ost_apps_2
OSS2	ost_home_3	ost_apps_3
	ost_home_4	ost_apps_4
	ost_home_5	ost_apps_5
OSS3	ost_home_6	ost_apps_6
	ost_home_7	ost_apps_7
	ost_home_8	ost_apps_8
OSS4	ost_home_9	ost_apps_9
	ost_home_10	ost_apps_10
	ost_home_11	ost_apps_11

Table 4: Summary of the OSTs mounted on the OSS servers

3.5.2 Multipath configuration for meta-data storage:

There are two virtual disks/volumes that are exported from the MDT controller to the MDSs. Each volume appears to the server as two separate disks. Therefore, they actually appear like four different devices. These are */dev/sda, /dev/sdb, /dev/sdc* and */dev/sdd.* Using these devices, two multipath devices are created. They are */dev/mapper/apps-MDS*(with */dev/sda* and */dev/sdc*) and */dev/mapper/home-MDS*(with */dev/sdb* and */dev/sdd*).

The meta-data storage servers do not use these 2 multipath block devices directly. Instead, Linux Volume Manager v2(LVM v2) is used to provide more flexibility. This is needed because the MDSs actually need 3 MDTs while we have only 2 block devices. Therefore, initially one LVM volume group is created per multipath device then logical volumes are created as required. Volume group named **vg_apps** is created on ***/dev/mapper/apps-MDS*** and **vg_home** on ***/dev/mapper/home-MDS***. In the volume group **vg_apps** only one logical volume ***/dev/mapper/vg_apps-mdt***(1.2TB in size) is created. In the volume group **vg_home** 2 logical volumes, ***/dev/mapper/vg_home-mgs***(25GB in size) and ***/dev/mapper/vg_home-mdt***(1.2TB in size) are created. The two LVM logical volumes **vg_home-mdt** & **vg_apps-**

mdt act as the MDTs for the respective file systems. One MGS volume **vg_home-mgs** is used for storing the internal configuration of the Lustre file systems. Refer Table 5 for details:

Volumes exported from EF3015 controller (show volumes)	Volumes as they appear to MDS (lsscsi)	Configured multipath device/alias (multipath – ll)	LVM Volume Group name (vgdisplay –s)	LVM Logical Volume name (lvdisplay)	Size
mdt-apps	/dev/sdc	dm-3(apps-MDS)	vg_apps	/dev/vg_apps/mdt	1.2TB
	/dev/sda				
mdt-home	/dev/sdd	dm-2(home-MDS)	vg_home	/dev/vg_home/mgs	25GB
	/dev/sdb			/dev/vg_home/mdt	1.2TB

Table 5: Volumes exported from the EF3015 Controller

4 Software Configuration

4.1 Storage Servers

Ivy Cluster uses Lustre in HA configuration with six storage servers on the whole, responsible for providing shared Lustre file system to all the compute nodes and head nodes of the Ivy Cluster. Out of these six, two are configured as MDSs and four as OSSs. Of the two MDS servers one is active and the other is in hot stand-by mode. In case of OSS, two OSSs i.e OSS1 and OSS2 work in active-active mode sharing the load on Lustre file system IOs. Same is the case with OSS3 and OSS4.

4.1.1 RAID support for Operating System

Each of the six storage servers i.e. MDS[1-2] and OSS[1-4] have two 300GB Hard disks configured as RAID 1 array in the software. This combined RAID array is used to install the OS on the storage server. This helps the server to withstand single disk failure scenario.

The RAID device name assigned to the software RAID array is ***/dev/md126***. To get more information about this device run mdadm as below:

```
# mdadm -D /dev/md126
```

On the RAID array, two partitions are created, one(***/dev/md126p1***) is for ***/boot*** and other(***/dev/md126p2***) is a LVM physical volume. On this LVM physical volume **VolumeGroup00** is created and the required logical volumes (like***/var*** and ***/***) for installing the Linux OS are created. On all the storage servers i.e. MDS[1-2] and OSS[1-4] CentOS 6.6 is installed.

4.1.2 Network Time Protocol(NTP)

The Lustre storage requires synchronized clocks for the storage system to work properly. Therefore the Ivy Cluster Head node is configured as NTP server and all the storage servers including the compute nodes are configured as NTP clients to synchronize their time with the Head nodes.

4.1.3 Access to LDAP servers

On Ivy cluster the user authentication is done by LDAP server. Therefore, the storage servers also should have access to the LDAP servers. The storage servers need access to the list of users who can access the cluster file-system. For this the head node of the Ivy cluster is set as the default gateway on all the storage servers. Also the head nodes are configured to **masquerade** the connections from **eth0(internal network 10.0.X.X of the cluster)** through **eth1(HPC network 10.20.X.X)** so that the LDAP requests coming from the storage servers and compute nodes are forwarded to the LDAP server.

4.2 EXAScalar Configuration file

4.2.1 MDS & OSS

As discussed earlier the Ivy Cluster Storage is configured with High Availability(HA) functionality. This configuration can be changed as required by modifying the '*/etc/ddn/exascaler.conf*' file.

4.2.2 Starting the storage (Lustre file system) resources

EXAScalar storage provides a python script named cluster_resources. This can be used to start or stop the Lustre file system services. It can be invoked on **mds1** as below:

```
# cluster_resources start

# cluster_resources stop
```

Once the resources are started the status can be found by running another python script named *hastatus* as below:

```
# hastatus
```

This will display the status of all the OSTs and MDTs, whether they are available or offline.

4.3 DDN SFA7700X Controller CLI (for OSTs)

The SFA Controller provides a command line interface and also a web interface to view or modify the configuration of the storage. The web interface can be accessed from a web browser on management node at **http://<IP of the storage controller>** like http://30.0.6.19/ for SFA1 primary controller. An ssh session can be started to the SFA controller from the management node of the Ivy Cluster as follows:

```
# ssh  user@30.0.6.19

# ssh  user@30.0.6.23
```

When asked for password enter **'user'** without quotes.

4.3.1 Listing the Physical Disks

The list of physical disks can be seen by issuing the following command:

```
# show phy
```

4.3.2 Viewing the enclosure status

The enclosure status can be seen by issuing the following command:

```
# show enc
```

4.3.3 Viewing the storage pools/Virtual Disks

The list of storage pools/Virtial Disks can be seen by issuing the following command:

```
# show pool
# show vd
```

4.3.4 Viewing the controller status

The controller status can be seen by issuing the following command:

```
# show con
```

4.4 DDN EF3015 Controller CLI (for MDTs)

The EF3015 Controller provides a command line interface and also a web interface to view or modify the configuration of the storage. Web interface can be accessed by the IP address of the controller similar to that of the SFA controller. An ssh session can be started to the EF3015 controller from the management node of the Ivy Cluster as follows:

```
# ssh  manage@30.0.6.27
```

When asked for password enter **'!manage'** without quotes.

4.4.1 Listing the Physical Disks

The list of physical disks can be seen by issuing the following command:

```
# show disks
```

4.4.2 Viewing the Virtual Disks

The list of virtual disks can be seen by issuing the following command:

```
# show vdisks
```

4.4.3 Viewing the volumes

The list of volumes exposed to the MDSs can be seen by issuing the following command:

```
# show volumes
```

5 Lustre File systems of Ivy Cluster

The EXAScaler(Lustre) file system is actually composed of two separate layers.

The lower layer of the file system is, at the OS level, an ext3 or ext4 type file system on each storage target. That is each MDT or OST is a local file system that can be mounted independently from the underlying storage. Within the storage system, there are three MDTs and twenty four OSTs as discussed earlier. Therefore same number of Ext3/Ext4 file systems exist with one per each target.

The upper layer of the file system is the Lustre file system formed by combining all the lower layer components i.e all MDTs and OSTs. This Lustre file system is mounted by compute nodes and master nodes and provides the parallel access. There are two Lustre file systems on the Ivy Cluster as listed in Table 6: Lustre file systems on Ivy Cluster.

Device	Mount on	Size	Purpose
20.0.6.13@o2ib:20.0.6.16@o2ib:/home on half of the compute nodes 20.0.6.14@o2ib1:20.0.6.17@o2ib1:/home on rest of the compute nodes	/home	256TB	For storing user home directories
20.0.6.13@o2ib:20.0.6.16@o2ib:/apps on half of the compute nodes 20.0.6.14@o2ib1:20.0.6.17@o2ib1:/apps on rest of the compute nodes	/apps	256TB	1. Cluster applications and 2. One scratch directory per user

Table 6: Lustre file systems on Ivy Cluster

5.1 Settings related to file striping pattern of Lustre

Files stored on the Lustre file systems(***/home*** and ***/apps***) are striped across multiple OSTs for achieving good performance and balanced usage of all the OSTs. For this there are few settings that can be modified per file basis or for all the files in a directory. This is done through the **lfs setstripe** Lustre command. The following three configurable parameters are assigned the default values as mentioned below:

stripe_size: The stripe size is the amount of data, of a given file, written to one OST before switching to the next OST. *The default stripe_size is 1 MB* for both ***/home*** and ***/apps*** file systems. While setting the stripe size of a file, passing a stripe_size of 0 causes the default stripe size to be used. The stripe_size value must be a multiple of 64 KB.

stripe_count: The stripe count indicates how many OSTs to use for the specified file. *The default stripe_count value is 1.* Setting stripe_count to 0 causes the default stripe count to be used. Setting stripe_count to -1 means stripe over all available OSTs (full OSTs will be skipped).

stripe_offset/start_ost_idx: The stripe offset is the index of the first OST to which the first stripe of a given file has to be written. *The default value for stripe_offset is -1*, which means the MDS decides the starting index. This setting allows space and load balancing to be done by the MDS as needed. If this setting is not -1, the file starts on the specified OST index. The numbering of the OSTs starts at 0.

These values can be changed by issuing the following command:

```
# lfs setstripe [--stripe-count|-c <stripe_count>] [--stripe-index|-i <start_ost_idx>] [--stripe-size|-S <stripe_size>] <directory|filename>
```

The following example sets the striping pattern of the files stored in */home/eig* directory such that the stripe size is 10MB, with contents striped across single OST and every new file gets stored on a different OST in round-robin fashion.

```
# lfs setstripe -S 10m -c 1 -i -1 /home/eig
```

6 Configuration of storage quota

Since all users of the Ivy Cluster have access to the same storage, there should be a mechanism to limit the usage of this common storage for each user, based on a policy. This is done by configuring the Storage Quota limits. This allows all the users to have their own share of the storage. If any user crosses his limit he has to delete the unwanted files to free up the space.

6.1 Quota for Lustre filesystem

Quota can be configured in Lustre filesystem using the **lfs** command. The quota settings consist of six separate fields.

1. Soft limit for block size
2. Hard limit for block size
3. Soft limit for inodes(no of files)
4. Hard limit for inodes
5. Block grace time
6. Inode grace time

The above six settings can be applied/changed for each file system separately. Like in the case of Ivy Cluster, we have '*/home*' and '*/apps*' which are two separate Lustre file systems. Therefore, the limits on '*/home*' can be different from limits on '*/apps*'.

Also out of the above mentioned limits the first four limits can be applied/changed per user basis or per group basis. But the last two i.e. the grace times are common for all the users within the file system.

6.1.1 Quota for Users

As an example, consider setting the quota for the users named 'cvsri' and 'radha'.

Setting user quota for cvsri user in '*/home*':

```
# lfs setquota -u cvsri -b 495G -B 500G /home/
```

Setting user quota for radha user in '*/apps*':

```
# lfs setquota -u radha -b 5000G -B 5T /apps/
```

To get the existing quota for a user radha in '*/home*':

```
# lfs quota -u radha /home/
```

6.1.2 Quota for Groups:

As an example, consider setting the quota for the groups named 'rseg' and 'cg'.

Setting group quota for rseg group in '*/apps*':

```
# lfs setquota -g rseg -b 20000G -B 20T /apps/
```

To get the existing quota for a group cg in '*/apps*':

```
# lfs quota -g cg /apps
```

6.1.3 Grace time:

Grace time is the time limit before the soft limit is enforced for a file system with quota enabled.

To change the block grace time for users in '*/home*':

```
# lfs setquota -t -u -b 99w /home/
```

To change the block grace time for groups in '*/apps*':

```
# lfs setquota -t -g -b 99w /apps/
```

To get the existing grace time:

```
# lfs quota -t -u /home
```

6.1.4 Qunit:

Qunit is the smallest unit of resource(inodes or blocks) allocated to the user at once. Limits that are not multiple of this qunit may actually be enforced after the qunit is crossed. For Example: If Qunit is 1MB then setting the hard limit to 1.5MB will cause the actual threshold to be 2MB.

6.2 Settings specific to Ivy Cluster storage:

Ivy Cluster contains two Lustre filesystems. These two filesystems are mounted on all the nodes of the Ivy Cluster as mentioned in section 5

Each user on the cluster will have two directories assigned to him on the '*/home*' and '*/apps*' filesystems. For ease of switching between these two directories, aliases are provided.

1. User home: **/home/<GroupName>/<UserName>** (type **uhm** at prompt to switch to this dir)
2. Scratch: **/apps/scratch/<GroupName>/<UserName>** (type **shm** at prompt to switch to this dir)

Quotas are configured for each user for the above two directories, so that every user gets a fair and equal storage space as below:

1. Maximum of 500GB on home
2. Maximum of 5TB on scratch.

Users who are crossing this limit have to delete unwanted files from their respective folder to copy any new content.

Apart from the user quota, group quota is also configured such that the space occupied by all the users in one group cannot cross 20TB on scratch(*'/apps'* filesystem). There is no such group quota restriction on *'/home'*. Therefore, they are free to use full 500GB under home directory

Quota settings	Soft limit	Hard limit	Grace time
/home user limit	495GB	500GB	99weeks
/apps user limit	5000GB	5TB	99weeks
/apps group limit	20000GB	20TB	99weeks

Table 7: Quota configuration limits on Ivy Cluster

6.3 Checking the quota from user perspective:

Users can check their quota by running the following commands:

On **home** filesystem: **$ lfs quota -h -u <username> /home** or by just **$ uquota**

On **scratch** filesystem: **$ lfs quota -h /apps** or by just typing **$ squota**

Users have to ensure that the **Used** space does **not cross** the **Limit**(i.e **500GB user limit on /home** and **5TB user limit or 20TB group limit on /apps).**

7 IP Addresses

The details about the IP Addresses assigned at the network interface level for various storage servers are given in Table 8.

Server	eth0	eth1	eth2	IPMI	ib0	ib1
oss1	10.0.6.1	30.0.6.1	30.0.6.2	30.0.6.3	20.0.6.1	20.0.6.2
oss2	10.0.6.2	30.0.6.4	30.0.6.5	30.0.6.6	20.0.6.4	20.0.6.5
oss3	10.0.6.3	30.0.6.7	30.0.6.8	30.0.6.9	20.0.6.7	20.0.6.8
oss4	10.0.6.4	30.0.6.10	30.0.6.11	30.0.6.12	20.0.6.10	20.0.6.11
mds1	10.0.6.5	30.0.6.13	30.0.6.14	30.0.6.15	20.0.6.13	20.0.6.14
mds2	10.0.6.6	30.0.6.16	30.0.6.17	30.0.6.18	20.0.6.16	20.0.6.17

Table 8: The storage servers and corresponding IP addresses

8 Summary

The high performance computing facility at Computer Division has multiple cluster computers for meeting the computational requirements of scientists and engineers of IGCAR. Ivy Cluster is one among them and this document gives the configuration details of Ivy Cluster storage system. The hardware and software configuration details of the storage system are covered in the document. The details about the Lustre file system are also covered.