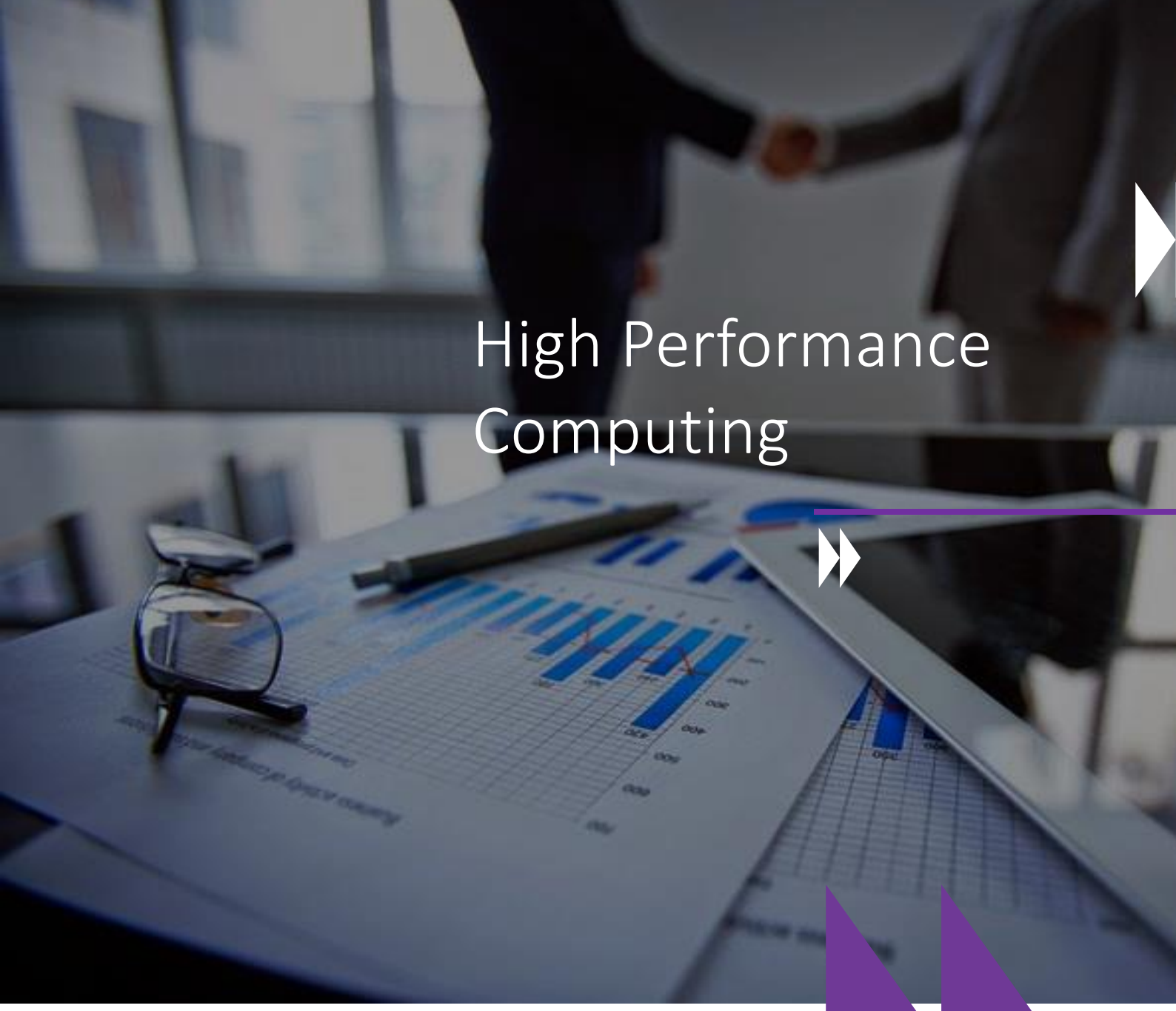


High Performance Computing

TIFR Hyderabad

Submitted to: TIFR Hyderabad

01-05-2019

Document Version and Confidentiality clause

Document Control

Name of the Document	HPC Installation Document
Client	TIFR Hyderabad
Reference	
Date	01-05-2019
Revision	Ver 1.0
Author	Sooraj Othayoth
Doc Type	Strictly Confidential

Table of Contents

About Us

Convergence is not just in our logo. Convergence is our credo. We started in 2000, with a belief that the world of Technology Infrastructure will converge. Now, we see convergence happening inside the datacenter of every enterprise. We are a trusted IT partner to many such businesses. We are helping them Converge to the Cloud. We are Locuz'

Locuz is an IT Infrastructure Solutions and Services Company focused on helping enterprises transform their businesses thru innovative and optimal use of technology. Our strong team of specialists, help address the challenge of deploying & managing complex IT Infrastructure in the face of rapid technological change. The greatest of these changes is the Cloud; and Locuz is uniquely positioned to help enterprises leverage the power of cloud technologies while avoiding the pitfalls of security, identity and service management.

- Locuz specializes in Hybrid Cloud Computing, Datacentre Transformation, NextGen Networking, Enterprise Collaboration, Information Security & Identity Management, High Performance Computing, Big Data & IT Automation
- Locuz serves many large and mid-size enterprises and institutions in varied segments such as Hi Tech, Life sciences, Healthcare, Financial Services, Insurance, Engineering Design, Education & Research

Apart from providing a wide range of advisory, implementation & managed IT services, Locuz has built innovative platforms in the area of Hybrid Cloud Orchestration, High Performance Computing & Software Asset Analytics. These products have been successfully deployed in leading enterprises in India and are helping customers extract greater RoI from their IT Infrastructure assets & investments.

ClouTor ganana aCube

Our Global Delivery centre in Chennai which houses our NOC & SOC, our Product Engineering & Innovation Centre in Bangalore and our 24/7 Service Desk and our field engineers spread across the country form the execution backbone of Locuz.

Headquartered in Hyderabad, India, Locuz has operations across all major cities in India. Locuz Inc is a wholly owned subsidiary of Locuz and is based in Texas, US. Locuz currently has over 350 people and serves over 200 customers

High availability xCAT Installation

Configuration Requirements

- Both the Primary and the Standby Management node should run the Identical Operating system
- xCAT version and database version should be identical on the two management nodes
- It is recommended to have similar hardware capability on the two management nodes to support the same operating system and have similar management capability.
- Static IP address should be assigned on both the management nodes
- Virtual IP Alias Address to be assigned on the Primary Management node
- The primary management node and standby management node should be in the same subnet to ensure the network services will work correctly after failover.
- Shared storage available for both the management node
- All the nodes should the BMC assigned with a Static IP
- All the nodes should have an identical User and Password for the BMC Network.

Management Node Operating system installation & Tuning

- Ensure a hostname is configured on the management node by issuing the hostname command. [*It's recommended to use a fully qualified domain name (FQDN) when setting the hostname*]
  1. To set the hostname of *xcatmn.cluster.com*:

     ```
     hostname master1.local
     ```

  2. Add the hostname to the `/etc/sysconfig/network` in order to persist the hostname on reboot.
  3. Reboot the server and verify the hostname by running the following commands:
     a. `hostname`
     b. `hostname -d` - should display the domain

- Assing the IP to **STATIC** in the `/etc/sysconfig/network-scripts/ifcfg-<dev>` configuration files.
- Configure any domain search strings and nameservers to the `/etc/resolv.conf` file.

Xcat installation

[root@master1 ~]# rpm -qa|grep -i xcat

grub2-xcat-2.02-0.16.el7.snap201506090204.noarch

syslinux-xcat-3.86-2.noarch

xCAT-buildkit-2.12.4-snap201611102359.noarch

conserver-xcat-8.1.16-10.x86_64

xCAT-genesis-base-x86_64-2.12-snap201607221057.noarch

xCAT-probe-2.12.4-snap201611102359.noarch

perl-xCAT-2.12.4-snap201611102359.noarch

xCAT-genesis-base-ppc64-2.12-snap201610250931.noarch

xCAT-2.12.4-snap201611102359.x86_64

xCAT-client-2.12.4-snap201611102359.noarch

xCAT-server-2.12.4-snap201611102359.noarch

xCAT-genesis-scripts-x86_64-2.12.4-snap201611102359.noarch

ipmitool-xcat-1.8.17-1.x86_64

xCAT-genesis-scripts-ppc64-2.12.4-snap201611102359.noarch

elilo-xcat-3.14-4.noarch

Xcat Version

[root@master1 ~]# lsxcatd -a

Version 2.12.4 (git commit 64ec2d41285d9a8770d8d9ef909f251ecbb5100b, built Thu Nov 10 23:59:02 EST 2016)

This is a Management Node

dbengine=SQLite

- Directory required to run the services in HA.

/install /etc/xcat /tftpboot /root/.xcat

4 services required to run xcat

xcatd dhcpd httpd named

To check xcat service working properly

[root@master1 ~]# tabdump site

#key,value,comments,disable

"blademaxp","64",,

"domain","local",,

"fsptimeout","0",,

"installdir","/install",,

"ipmimaxp","64",,

"ipmiretries","3",,

"ipmitimeout","2",,

"consoleondemand","no",,

"master","192.168.102.6",,

"forwarders","192.168.102.6",,

"nameservers","192.168.102.6",,

"ntpserver","192.168.102.6",,

"maxssh","8",,

"ppcmaxp","64",,

"ppcretry","3",,

"ppctimeout","0",,

"powerinterval","0",,

"syspowerinterval","0",,

"sharedtftp","1",,

"SNsyncfiledir","/var/xcat/syncfiles",,

"nodesyncfiledir","/var/xcat/node/syncfiles",,

"tftpdir","/tftpboot",,

"xcatdport","3001",,

"xcatiport","3002",,

"xcatconfdir","/etc/xcat",,

"timezone","Asia/Kolkata",,

"useNmapfromMN","no",,

"enableASMI","no",,

"db2installloc","/mntdb2",,

"databaseloc","/var/lib",,

"sshbetweennodes","ALLGROUPS",,

"dnshandler","ddns",,

"vsftp","n",,

"cleanupxcatpost","no",,

"dhcplease","43200",,

"auditnosyslog","0",,

"xcatsslversion","TLSv1",,

"auditskipcmds","ALL",,

"mnroutenames","defaultroute",,

"dhcpinterfaces","eno1",,

To check node details

#lsdef <nodename>

[root@master1 ~]# lsdef node01

Object name: node01

arch=x86_64

bmc=192.168.102.11

bmcpassword=superuser

bmcusername=root

cons=ipmi

currchain=boot

currstate=boot

groups=compute,all

initrd=xcat/osimage/centos7.6-x86_64-install-compute/initrd.img

installnic=mac

ip=192.168.103.11

kcmdline=quiet inst.repo=http://!myipfn!:80/install/centos7.6/x86_64 inst.ks=http://!myipfn!:80/install/autoinst/node01 ip=dhcp

kernel=xcat/osimage/centos7.6-x86_64-install-compute/vmlinuz

mac=A4:BF:01:2E:9C:1D

mgt=ipmi

netboot=xnba

nicips.eno1=192.168.103.11

nicnetworks.eno1=192_168_102_0-255_255_254_0

nictypes.eno1=ethernet

os=centos7.6

postbootscripts=otherpkgs

postscripts=syslog,remoteshell,syncfiles,confGang,configNetwork,setroute replace,setroute add

primarynic=mac

profile=compute

provmethod=centos7.6-x86_64-install-compute

routenames=defaultroute

status=booted

statustime=04-09-2019 15:31:51

updatestatus=synced

updatestatustime=03-31-2019 02:18:32

To add new node in cluster.

e.g to add node01

#chdef -t node node01 groups=compute,all mgt=ipmi cons=ipmi ip=192.168.103.11 netboot=xnba bmc=192.168.102.11 bmcusername=root bmcpassword=superuser installnic=mac primarynic=mac mac=A4:BF:01:2E:9C:1D

#makehosts

#makedhcp node01

#makedns node01

#nodeset node01 osimage=centos7.6-x86_64-install-compute

Boot the nodes..

To remove the nodes from cluster

#rmdef node01

remove the host entry from /etc/hosts

#makehosts

To reinstall nodes.

#rinstall node01

Power off the nodes

#rpower node01 off

Power on the nodes

#rpower node01 on

Power reset the nodes

#rpower node01 reset

To run command parallel

#xdsh node[01-88] "date"

Configure the Base OS Repository

xCAT uses the yum package manager on RHEL Linux distributions to install and resolve dependency packages provided by the base operating system.

1. Mount the iso to `/repo` on the Management Node.

```
mkdir -p /repo
mount -o loop /tmp/RHEL-LE-7.6 -Server-ppc64le-dvd1.iso /repo
```

2. Create a yum repository file `/etc/yum.repos.d/local.repo`

```
[local]
name=local
baseurl=file:///repo
gpgcheck=0
enabled=1
```

1. Mount the shared storage available on the primary management node.

```
The Compute Storage is mounted on Primary Management node on the following
filesystem path

192.168.12.3:/xcat_ha/install     195G   84G  111G  44% /install
192.168.12.3:/xcat_ha/etc/xcat    195G   84G  111G  44% /etc/xcat
192.168.12.3:/xcat_ha/root/.xcat  195G   84G  111G  44% /root/.xcat
192.168.12.3:/xcat_ha/tftpboot    195G   84G  111G  44% /tftpboot
```

2. Set up a "Virtual IP address" that will be floating across the Primary & Standby Management nodes during failover.

Install Xcat

```
#yum clean all (optional)
#yum install xCAT
```

3. Disable xcat managed services to start automatically.

```
#service xcatd stop
#chkconfig --level 345 xcatd off
#service conserver off
#chkconfig --level 2345 conserver off
#service dhcpd stop
#chkconfig --level 2345 dhcpd off
```

Note: During the install, you must accept the *xCAT Security Key* to continue

Verify xCAT

```
#source /etc/profile.d/xcat.sh
"#lsxcatd -a" should print the xCAT version
"#tabdump site" will dumpout the site table.
```

The output should be similar to the following:

```
#key,value,comments,disable
"blademaxp","64",,
"domain","pok.stglabs.ibm.com",,
"fsptimeout","0",,
"installdir","/install",,
"ipmimaxp","64",,
"ipmiretries","3",,
...
```

- To start xCAT

```
#systemctl start xcatd.service
```

- To stop xCAT

```
#systemctl stop xcatd.service
```

- To restart xCAT

```
#systemctl restart xcatd.service
```

- To verify xCAT service

```
#systemctl status xcatd.service
```

xCAT configuration

NOTE: *Mount the shared storage to the node before starting to configure xCAT, The following configurations should be done on the Primary management node.*

XCAT use 'copycds' command to create an image which will be available to install nodes. "copycds" will copy all contents of Distribution DVDs/ISOs or Service Pack DVDs/ISOs to a destination directory, and create several relevant osimage definitions by default.

1. Import the Compute node OS Distribution to xCAT

```
#copycds /software/rhel-server-7.6-x86_64-dvd.iso
```

2. Verify the list of OS images created in xCAT

```
#lsdef -t osimage
```

3. Below is an example of osimage definitions created by copycds:

```
# lsdef -t osimage
centos7.6-x86_64-install-compute (osimage)
centos7.6-x86_64-install-service (osimage)
centos7.6-x86_64-netboot-compute (osimage)
centos7.6-x86_64-stateful-mgmtnode (osimage)
```

4. Configure the DNS table by executing the following command.

```
#makedns  -n
#systemctl restart named
```

5. Verify the node attributes as follows

```
# lsdef -t node node01
```

6. Preparing the Stateless Osimage

Copy the osimage for stateless with the customized image name as follows

```
#lsdef -t osimage -z rhels7.6-x86_64-netboot-compute | sed 's/^[^
]\+:/iafcomputeimage:/' | mkdef -z
```

7. Cofigure the Dynamic IP address Range which will be assigned to the compute node during pxeboot

```
#chdef -t network 192.168.103_0-255_255_254_0
dynamicrange="192.168.103.11-192.168.103.254
```

8. Create the initrd image for the diskless boot as follows

```
#genimage iafcomputeimage
```

9. create the compressed root image for the diskless boot as follows,

```
#packimage iafcomputeimage

#mknb x86_64
```

10. Add the compute nodes as follows

```
# mkdef node01 groups=compute ip=192.168.103.11 netboot=xnba
installnic=mac primarynic=mac mac=44:37:e6:16:1c:0a
```

follow the above command to add all the compute node by substituting the “nodename”,ip=”ip address for the compute node”,mac=”mac address of the compute node”

11. Update the xCAT DB by executing the following commands

```
#makenetworks

#makehosts

#makedhcp -a

#makedns -a
```

NOTE: the ip address used in the above command should not fall in the dynamic range which is configured in the previous step (step7)

```
#nodeset “nodename” osimage=centos7.2-x86_64-netboot-compute
```

Execute the above command by substituting the “nodename” as compute node hostname for all the compute nodes in the cluster.

Configuring PCSD for High availability on the Management node

1. Disable xCAT managed services to start automatically

```
#service xcatd stop
#chkconfig --level 345 xcatd off
#service conserver off
#chkconfig --level 2345 conserver off
#service dhcpd stop
#chkconfig --level 2345 dhcpd off
```

2. Mount the shared data on all the mountpoints (/install, /etc/xcat & /root/.xcat)

3. Installing pacemaker and add-on tool and start pcs

#yum install pcs pacemaker fence-agents-all

#passwd hacluster

#systemctl start pcsd

#systemctl status pcsd

#systemctl enable pcsd

4. create a cluster by adding both the management nodes

pcs cluster auth xcatmanagmentnode1 xcatmanagementnode2

#Username: hacluster

#Password:

#xcatmanagmentnode1: Authorized

#xcatmanagementnode2: Authorized

NOTE: when prompted enter the password which is created in the previous step (step 2)

#pcs cluster setup --start --name HPC_XCAT xcatmangementnode1 xcatmanagementnode2

The above command will create the cluster and sync the configuration data on the cluster members and the output of the command should be as follows,

>Sending cluster config files to the nodes...

>xcatmangementnode1: Succeeded

>xcatmangementnode2: Succeeded

>Starting cluster on nodes: xcat1, xcat2...

>xcatmangementnode1: Starting Cluster...

>xcatmangementnode2: Starting Cluster...

>Synchronizing pcsd certificates on nodes xcat1, xcat2...

>xcatmangementnode1: Success

>xcatmangementnode2: Success

>Restarting pcsd on the nodes in order to reload the >certificates...

>xcatmangementnode1: Success

>xcatmangementnode2: Success

#pcs cluster enable –all

>xcat1: Cluster Enabled

>xcat2: Cluster Enabled

5. Creating Resource group and adding the resource to float across the primary and the standby Management node

#pcs resource create intVirtualIP IPaddr2 ip=10.0.0.100 cidr_netmask=24 --group xcatHA

#pcs resource create extVirtualIP IPaddr2 ip=10.129.150.100 cidr_netmask=24 --group xcatHA

#pcs resource create installshare Filesystem device="/dev/sdb1" directory="/install" fstype="xfs" --group xcatHA

#pcs resource create etcshare Filesystem device="/dev/sdb2" directory="/etc/xcat" fstype="xfs" --group xcatHA

#pcs resource create locuzlun Filesystem device="/dev/sdb3" directory="/root/.xcat" fstype="xfs" --group xcatHA

#pcs resource create XCAT service:heartbeat:dhcpd --group xcatHA –force

```
#pcs resource create xcatservice lsb:xcatd –group xcatHA
```

6. Fencing configuration for IPMI for STONITH config .

#pcs stonith create fence_xcatmgnt1 fence_ipmilan pcmk_host_list="xcatmgnt1" ipaddr="10.0.0.199" action="reboot" login="ADMIN" passwd="ADMIN" delay=15 op monitor interval=60s

#pcs stonith create fence_xcatmgnt2 fence_ipmilan pcmk_host_list="xcatmgnt2" ipaddr="10.0.0.250" action="reboot" login="ADMIN" passwd="ADMIN" delay=15 op monitor interval=60

7. Verify the status of all the resources by the following commands

#pcs status
#pcs cluster status
#pcs resource show

Master1 Node Software Configuration

Operating System	CentOS 7.6
Cluster Manager	xCat
Scheduler	PBS Pro
Monitoring Tool	Ganglia
Master1 Node	
Host Name	Master1.local
IP Address	Eth0- 192.168.102.1 Eth1- 10.10.11.16 IPMI - 192.168.102.2 IB - 192.168.12.1
	raid 1 with 960 GB SSD approx. usable space
Partition	/boot – 1 GB /var – 200 GB /tmp – 100GB / - 166 GB Swap –32 GB

Master2 Node Software Configuration

Operating System	CentOS 7.6
Cluster Manager	xCat
Scheduler	PBS Pro
Monitoring Tool	Ganglia
Master2 Node	
Host Name	Master2.local
IP Address	Eth0- 192.168.102.3 Eth1- 10.10.11.17 IPMI - 192.168.102.4 IB - 192.168.12.2
	raid 1 with 960 GB SSD approx. usable space
Partition	/boot – 1 GB /var – 200 GB /tmp – 100GB / - 166 GB Swap –32 GB

Ifconfig (Master1.local)

```
[root@master1 ~]# ifconfig
eno1: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.102.1  netmask 255.255.254.0  broadcast 192.168.103.255
        inet6 fe80::a6bf:1ff:fe67:8f03  prefixlen 64  scopeid 0x20<link>
        ether a4:bf:01:67:8f:03  txqueuelen 1000  (Ethernet)
        RX packets 22741602  bytes 6009769675 (5.5 GiB)
        RX errors 0  dropped 25  overruns 0  frame 0
        TX packets 9301805  bytes 862464113 (822.5 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
eno2: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 10.10.11.16  netmask 255.255.255.0  broadcast 10.10.11.255
        inet6 fe80::a6bf:1ff:fe67:8f04  prefixlen 64  scopeid 0x20<link>
        ether a4:bf:01:67:8f:04  txqueuelen 1000  (Ethernet)
        RX packets 1654690  bytes 273630907 (260.9 MiB)
        RX errors 0  dropped 11651  overruns 0  frame 0
        TX packets 147718  bytes 18863733 (17.9 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
ib0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 65520
        inet 192.168.12.1  netmask 255.255.255.0  broadcast 192.168.12.255
        inet6 fe80::211:7509:104:637a  prefixlen 64  scopeid 0x20<link>
Infiniband hardware address can be incorrect! Please read BUGS section in ifconfig(8).
infiniband 80:81:00:02:FE:80:00:00:00:00:00:00:00:00:00:00:00:00:00:00  txqueuelen 256  (InfiniBand)
        RX packets 2063458  bytes 771426329 (735.6 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 3221820  bytes 5800691591 (5.4 GiB)
        TX errors 0  dropped 47 overruns 0  carrier 0  collisions 0
```

df –h (master1.local)

```
[root@master1 ~]# df -h
Filesystem                      Size  Used Avail Use% Mounted on
/dev/md126p4                    166G   99G   67G  60% /
devtmpfs                         32G     0   32G   0% /dev
tmpfs                            32G   38M   32G   1% /dev/shm
tmpfs                            32G   28M   32G   1% /run
tmpfs                            32G     0   32G   0% /sys/fs/cgroup
/dev/md126p5                    100G   33M  100G   1% /tmp
/dev/md126p2                   1014M  184M  831M  19% /boot
/dev/md126p7                    350G   18G  333G   5% /apps
/dev/md126p1                   1022M   12M 1011M   2% /boot/efi
/dev/md126p3                    200G  3.2G  197G   2% /var
tmpfs                           6.3G   12K  6.3G   1% /run/user/42
192.168.12.3:/HA/PBS-HA         195G   94G  102G  48% /var/spool/pbs
192.168.12.3:/xcat_ha/install   195G   94G  102G  48% /install
192.168.12.3:/xcat_ha/etc/xcat  195G   94G  102G  48% /etc/xcat
192.168.12.3:/xcat_ha/root/.xcat 195G  94G  102G  48% /root/.xcat
nis192.168.12.3:/xcat_ha/tftpboot 195G 94G  102G  48% /tftpboot
tmpfs                           6.3G     0  6.3G   0% /run/user/0
192.168.12.3:/home-storage/home 190T  648G  190T   1% /home
```

/etc/fstab (master1.local)

```
[root@master1 ~]# cat /etc/fstab
UUID=4138f9e6-4998-4fea-89ec-0593f7f8bcb6 /       xfs     defaults                0 0
UUID=83668b94-8056-40c9-bc6b-20d3a9f7e65c /apps   xfs     defaults                0 0
UUID=a562567f-5ee8-482e-acec-e60769e53152 /boot   xfs     defaults                0 0
UUID=15F3-DA36                            /boot/efi vfat  umask=0077,shortname=winnt 0 0
UUID=c55e9f2c-d288-4302-b554-951a58b8b8fc /tmp    xfs     defaults                0 0
UUID=272490eb-89dc-4041-9b2d-19b696432eee /var    xfs     defaults                0 0
```

UUID=799ac88b-35a1-4ee4-b494-9ea366d9de2c swap swap defaults 0 0

#192.168.102.10:/xcat_ha/etc/xcat /etc/xcat nfs defaults,_netdev 0 0

#192.168.102.10:/xcat_ha/install /install nfs defaults,_netdev 0 0

#192.168.102.10:/xcat_ha/tftpboot /tftpboot nfs defaults,_netdev 0 0

#192.168.102.10:/xcat_ha/root/.xcat /root/.xcat nfs defaults,_netdev 0 0

Compute Storage

Storage Node	
Host Name	Storage.local
IP Address	Eth0- 192.168.102.10 IPMI - 192.168.102.9 IB - 192.168.12.3
	raid 1 with 240 GB SSD approx. usable space
Partition	/boot – 1 GB / - 195 GB Swap –16 GB /home/storage – 192TB

df –h (storage.local)

[root@storage ~]# df -h

Filesystem	Size	Used	Avail	Use%	Mounted on
/dev/md126p3	195G	94G	102G	48%	/
devtmpfs	126G	0	126G	0%	/dev
tmpfs	126G	0	126G	0%	/dev/shm
tmpfs	126G	20M	126G	1%	/run
tmpfs	126G	0	126G	0%	/sys/fs/cgroup
/dev/md126p2	1014M	185M	830M	19%	/boot
/dev/md126p1	1022M	12M	1011M	2%	/boot/efi
home-storage	192T	2.1T	190T	2%	/home-storage
tmpfs	26G	12K	26G	1%	/run/user/42
tmpfs	26G	0	26G	0%	/run/user/0
home-storage/home	190T	648G	190T	1%	/home-storage/home
localhost:/home-storage/home	190T	648G	190T	1%	/home

/etc/fstab (storage.local)

[root@storage ~]# cat /etc/fstab

UUID=506c327e-c6f6-48af-9128-24cd6401fc7e / xfs defaults 0 0

UUID=ddb643d3-b81a-4e99-a8bc-d902b1a97794 /boot xfs defaults 0 0

UUID=BC5D-FEBA /boot/efi vfat umask=0077,shortname=winnt 0 0

UUID=49f9c8a9-8812-4cd3-bfa8-3f3a386f0de9 swap swap defaults 0 0

Compute Node (node01-88)

Operating System	CentOS 7.6
Compute Node	
Host Name	Node[01-88]
IP Address	Eth0- 192.168.103.11 to 192.168.103.98 IPMI - 192.168.102.11 to 192.168.102.98 IB - 192.168.12.11 to 192.168.12.98
	240 GB SSD approx. usable space
Partition	/boot – 1 GB swap – 16 GB / - 207 GB

df –h (node00.local)

[root@node01 ~]# df -h

Filesystem Size Used Avail Use% Mounted on

/dev/sda3 207G 42G 166G 21% /

devtmpfs 47G 0 47G 0% /dev

tmpfs 47G 68M 47G 1% /dev/shm

tmpfs 47G 11M 47G 1% /run

tmpfs 47G 0 47G 0% /sys/fs/cgroup

/dev/sda1 1014M 162M 853M 16% /boot

192.168.12.3:/home-storage/home 190T 648G 190T 1% /home

tmpfs 9.3G 0 9.3G 0% /run/user/0

/etc/fstab (node00.local)

```
[root@node01 ~]# cat /etc/fstab
UUID=8badcdfe-5bd1-4adb-aac8-9e145bc5f97a /                       xfs     defaults        0 0
UUID=5fbeebd3-427e-4a47-9ea3-a7cb9b2042da /boot                   xfs     defaults        0 0
UUID=69dd443f-06f2-4357-9f55-102208705f51 swap                    swap    defaults        0 0
```

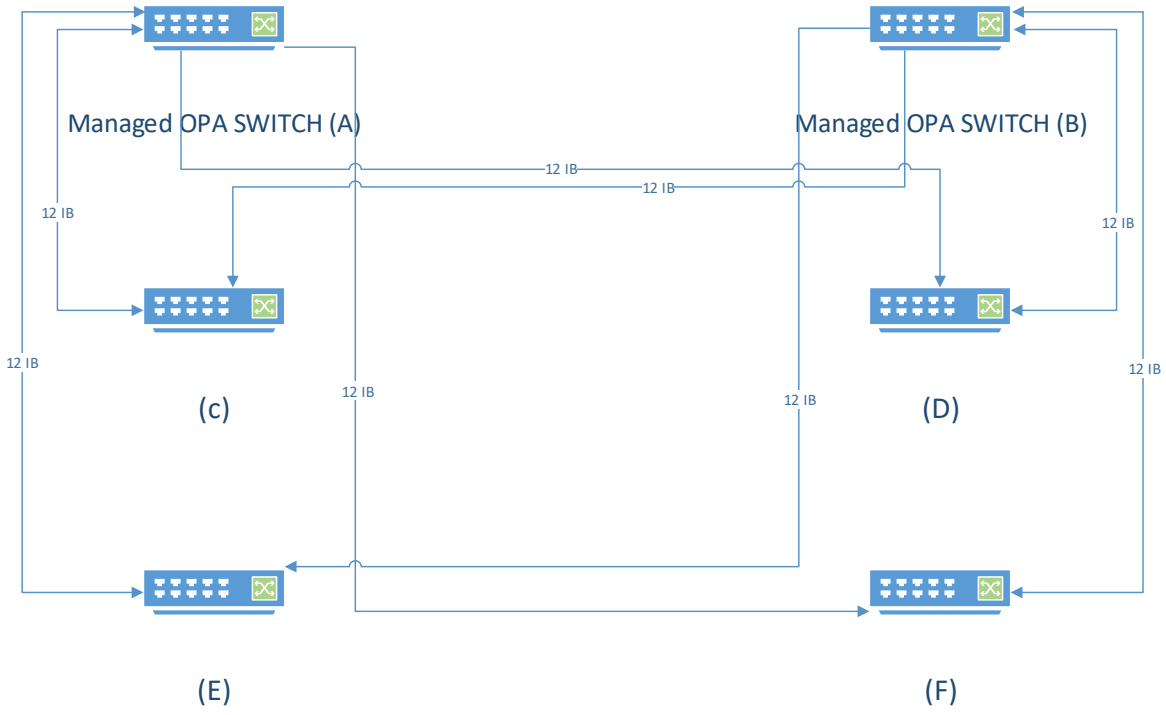
IP Address Details

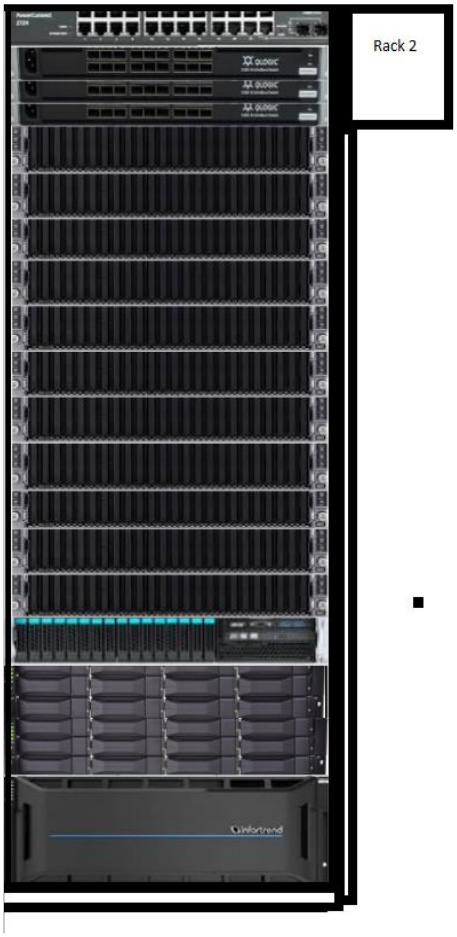
Node Name	IPMI IP Address	ETH IP Address	IB IP ADDRESS	Internal
Master Node1	192.168.102.3	192.168.102.1	192.168.12.1	10.10.11.16
Master Node 2	192.168.102.4	192.168.102.2	192.168.12.2	10.10.11.17
Compute Storage	192.168.102.9	192.168.102.10	192.168.12.3	
IB Switch 1	192.168.102.8			
IB Switch 2	192.168.102.7			
NODE NAME	**IPMI IPADDRESS**	**ETH IP ADDTESS**	**MAC ADDRESS**	**IB IP address**
node01	192.168.102.11	192.168.103.11	A4:BF:01:2E:9C:1D	192.168.12.11
node02	192.168.102.12	192.168.103.12	A4:BF:01:2E:97:13	192.168.12.12
node03	192.168.102.13	192.168.103.13	A4:BF:01:2E:96:D7	192.168.12.13
node04	192.168.102.14	192.168.103.14	A4:BF:01:2E:AC:44	192.168.12.14
node05	192.168.102.15	192.168.103.15	A4:BF:01:2E:03:1C	192.168.12.15
node06	192.168.102.16	192.168.103.16	A4:BF:01:2E:9C:4F	192.168.12.16
node07	192.168.102.17	192.168.103.17	A4:BF:01:2E:95:E2	192.168.12.17
node08	192.168.102.18	192.168.103.18	A4:BF:01:2E:9A:A1	192.168.12.18
node09	192.168.102.19	192.168.103.19	A4:BF:01:2E:08:62	192.168.12.19
node10	192.168.102.20	192.168.103.20	A4:BF:01:2D:EF:BE	192.168.12.20
node11	192.168.102.21	192.168.103.21	A4:BF:01:2E:99:5C	192.168.12.21
node12	192.168.102.22	192.168.103.22	A4:BF:01:2E:0A:5B	192.168.12.22
node13	192.168.102.23	192.168.103.23	A4:BF:01:2E:9C:C7	192.168.12.23
node14	192.168.102.24	192.168.103.24	A4:BF:01:2E:AE:29	192.168.12.24
node15	192.168.102.25	192.168.103.25	A4:BF:01:2E:96::46	192.168.12.25
node16	192.168.102.26	192.168.103.26	A4:BF:01:2E:AA:91	192.168.12.26
node17	192.168.102.27	192.168.103.27	A4:BF:01:2E:94:D9	192.168.12.27
node18	192.168.102.28	192.168.103.28	A4:BF:01:2E:98:21	192.168.12.28
node19	192.168.102.29	192.168.103.29	A4:BF:01:29:B2:67	192.168.12.29
node20	192.168.102.30	192.168.103.30	A4:BF:01:2E:EF:CD	192.168.12.30
node21	192.168.102.31	192.168.103.31	A4:BF:01:2E:0B:8C	192.168.12.31
node22	192.168.102.32	192.168.103.32	A4:BF:01:2E:AA:F0	192.168.12.32
node23	192.168.102.33	192.168.103.33	A4:BF:01:2E:AE:1F	192.168.12.33
node24	192.168.102.34	192.168.103.34	A4:BF:01:2E:AD:B6	192.168.12.34

node25	192.168.102.35	192.168.103.35	A4:BF:01:2E:97:B3	192.168.12.35
node26	192.168.102.36	192.168.103.36	A4:BF:01:2E:05:79	192.168.12.36
node27	192.168.102.37	192.168.103.37	A4:BF:01:2E:98:62	192.168.12.37
node28	192.168.102.38	192.168.103.38	A4:BF:01:2D:EF:9B	192.168.12.38
node29	192.168.102.39	192.168.103.39	A4:BF:01:2E:A7:1C	192.168.12.39
node30	192.168.102.40	192.168.103.40	A4:BF:01:2E:9D:FD	192.168.12.40
node31	192.168.102.41	192.168.103.41	A4:BF:01:2E:A8:D9	192.168.12.41
node32	192.168.102.42	192.168.103.42	A4:BF:01:2E:A6:5E	192.168.12.42
node33	192.168.102.43	192.168.103.43	A4:BF:01:2E:97:A9	192.168.12.43
node34	192.168.102.44	192.168.103.44	A4:BF:01:2D:E8:8E	192.168.12.44
node35	192.168.102.45	192.168.103.45	A4:BF:01:2E:0A:2E	192.168.12.45
node36	192.168.102.46	192.168.103.46	A4:BF:01:2E:05:8D	192.168.12.46
node37	192.168.102.47	192.168.103.47	A4:BF:01:2E:07:8B	192.168.12.47
node38	192.168.102.48	192.168.103.48	A4:BF:01:2E:06:05	192.168.12.48
node39	192.168.102.49	192.168.103.49	A4:BF:01:2E:08:A3	192.168.12.49
node40	192.168.102.50	192.168.103.50	A4:BF:01:2D:EB:FE	192.168.12.50
node41	192.168.102.51	192.168.103.51	A4:BF:01:2E:98:49	192.168.12.51
node42	192.168.102.52	192.168.103.52	A4:BF:01:2E:04:CF	192.168.12.52
node43	192.168.102.53	192.168.103.53	A4:BF:01:2E:06:0A	192.168.12.53
node44	192.168.102.54	192.168.103.54	A4:BF:01:2E:08:12	192.168.12.54
node45	192.168.102.55	192.168.103.55	A4:BF:01:2E:07:A9	192.168.12.55
node46	192.168.102.56	192.168.103.56	A4:BF:01:2E:0A:0B	192.168.12.56
node47	192.168.102.57	192.168.103.57	A4:BF:01:2E:97:B8	192.168.12.57
node48	192.168.102.58	192.168.103.58	A4:BF:01:2E:95:56	192.168.12.58
node49	192.168.102.59	192.168.103.59	A4:BF:01:2E:92:54	192.168.12.59
node50	192.168.102.60	192.168.103.60	A4:BF:01:2E:9B:46	192.168.12.60
node51	192.168.102.61	192.168.103.61	A4:BF:01:2E:96:FF	192.168.12.61
node52	192.168.102.62	192.168.103.62	A4:BF:01:2E:04:E3	192.168.12.62
node53	192.168.102.63	192.168.103.63	A4:BF:01:2E:AB:7C	192.168.12.63
node54	192.168.102.64	192.168.103.64	A4:BF:01:2E:9C:18	192.168.12.64
node55	192.168.102.65	192.168.103.65	A4:BF:01:2E:96:F5	192.168.12.65
node56	192.168.102.66	192.168.103.66	A4:BF:01:2E:AA:73	192.168.12.66
node57	192.168.102.67	192.168.103.67	A4:BF:01:2E:97:D1	192.168.12.67
node58	192.168.102.68	192.168.103.68	A4:BF:01:2E:96:6E	192.168.12.68
node59	192.168.102.69	192.168.103.69	A4:BF:01:29:92:50	192.168.12.69
node60	192.168.102.70	192.168.103.70	A4:BF:01:2E:AD:A2	192.168.12.70
node61	192.168.102.71	192.168.103.71	A4:BF:01:2E:9B:3C	192.168.12.71
node62	192.168.102.72	192.168.103.72	A4:BF:01:2E:97:DB	192.168.12.72
node63	192.168.102.73	192.168.103.73	A4:BF:01:2E:97:6D	192.168.12.73
node64	192.168.102.74	192.168.103.74	A4:BF:01:2E:AA:82	192.168.12.74
node65	192.168.102.75	192.168.103.75	A4:BF:01:2E:08:08	192.168.12.75
node66	192.168.102.76	192.168.103.76	A4:BF:01:2E:AB:C2	192.168.12.76
node67	192.168.102.77	192.168.103.77	A4:BF:01:2E:AC:B2	192.168.12.77
node68	192.168.102.78	192.168.103.78	A4:BF:01:2E:AA:DC	192.168.12.78
node69	192.168.102.79	192.168.103.79	A4:BF:01:2E:08:D0	192.168.12.79
node70	192.168.102.80	192.168.103.80	A4:BF:01:2E:00:FB	192.168.12.80

node71	192.168.102.81	192.168.103.81	A4:BF:01:2E:06:A5	192.168.12.81
node72	192.168.102.82	192.168.103.82	A4:BF:01:2E:A8:3E	192.168.12.82
node73	192.168.102.83	192.168.103.83	A4:BF:01:2E:08:B7	192.168.12.83
node74	192.168.102.84	192.168.103.84	A4:BF:01:2E:0A:FB	192.168.12.84
node75	192.168.102.85	192.168.103.85	A4:BF:01:2E:05:BF	192.168.12.85
node76	192.168.102.86	192.168.103.86	A4:BF:01:2E:01:41	192.168.12.86
node77	192.168.102.87	192.168.103.87	A4:BF:01:2E:A2:0D	192.168.12.87
node78	192.168.102.88	192.168.103.88	A4:BF:01:2E:AD:F2	192.168.12.88
node79	192.168.102.89	192.168.103.89	A4:BF:01:2E:0A:01	192.168.12.89
node80	192.168.102.90	192.168.103.90	A4:BF:01:2E:A5:91	192.168.12.90
node81	192.168.102.91	192.168.103.91	A4:BF:01:2E:AE:79	192.168.12.91
node82	192.168.102.92	192.168.103.92	A4:BF:01:2E:85:61	192.168.12.92
node83	192.168.102.93	192.168.103.93	A4:BF:01:2E:0B:D2	192.168.12.93
node84	192.168.102.94	192.168.103.94	A4:BF:01:2E:84:94	192.168.12.94
node85	192.168.102.95	192.168.103.95	A4:BF:01:2E:08:99	192.168.12.95
node86	192.168.102.96	192.168.103.96	A4:BF:01:2E:07:95	192.168.12.96
node87	**192.168.102.97**	192.168.103.97	**A4:BF:01:2E:09:02**	**192.168.12.97**
node88	192.168.102.98	192.168.103.98	A4:BF:01:2E:08:CB	192.168.12.98

InfiniBand OPA Switch connectivity (Factory Topology)

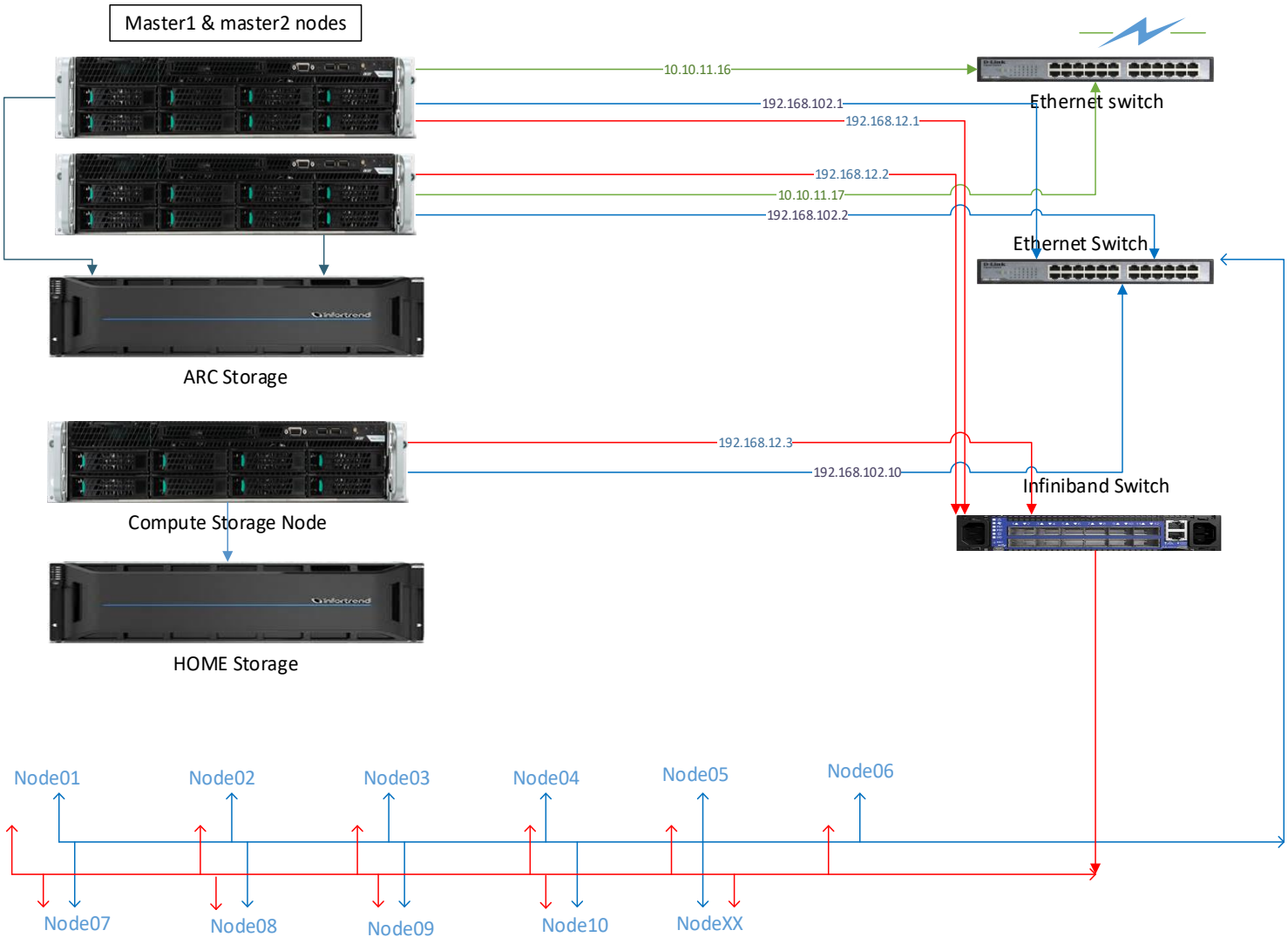




RACK VIEW DIAGRAM

COMPONENT	OEM	RACK UNIT SIZE	NOS
1)ETHERNET SWITCH	DELL	1U	2
2) COMPUTE NODES (Chassis)	ACER	2U	22
3)MASTER NODE	ACER	2U	2
4) JBOD	INFORTEND	4U	2
5) INFINIBAND SWITCH	INTEL	2U	6
6) KVM	ATEN	1U	1
7) RACK CABINET	RITTAL	42 U	2

Logical Diagram



IB Connectivity

Internet Connection

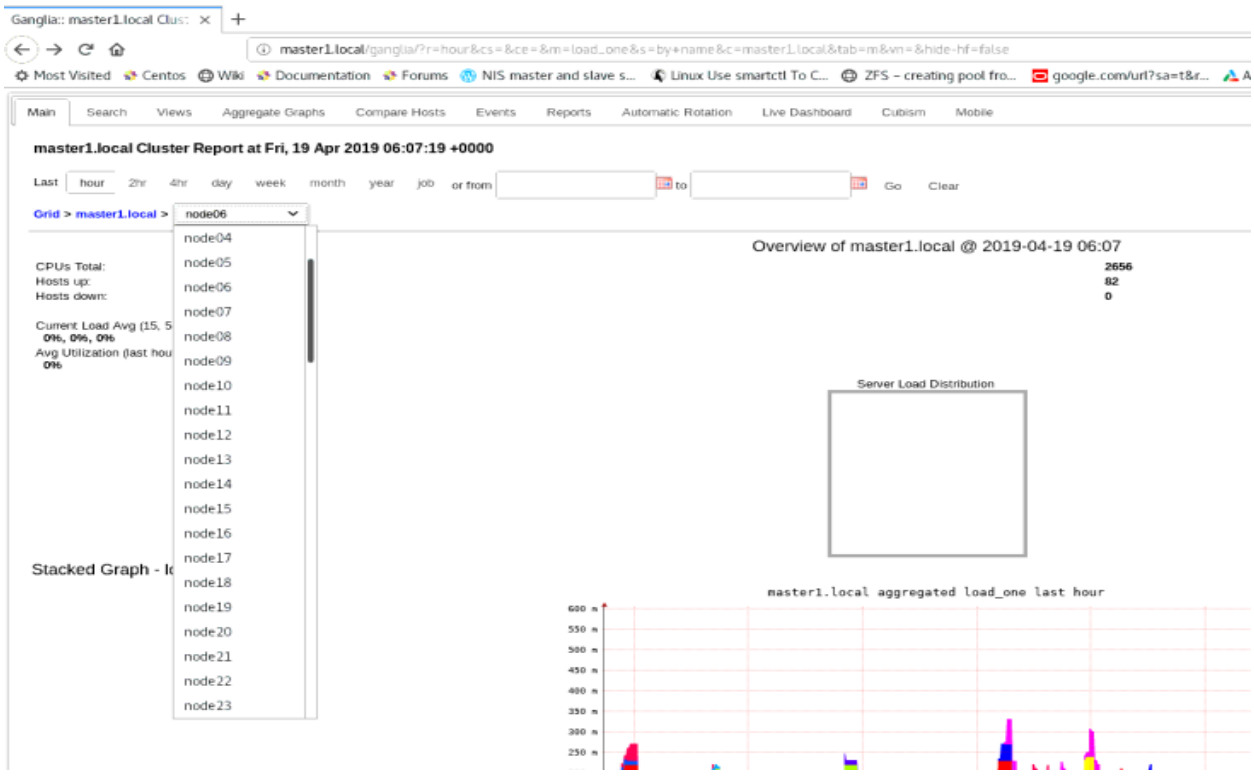
Inter Connection

Infiniband Connection

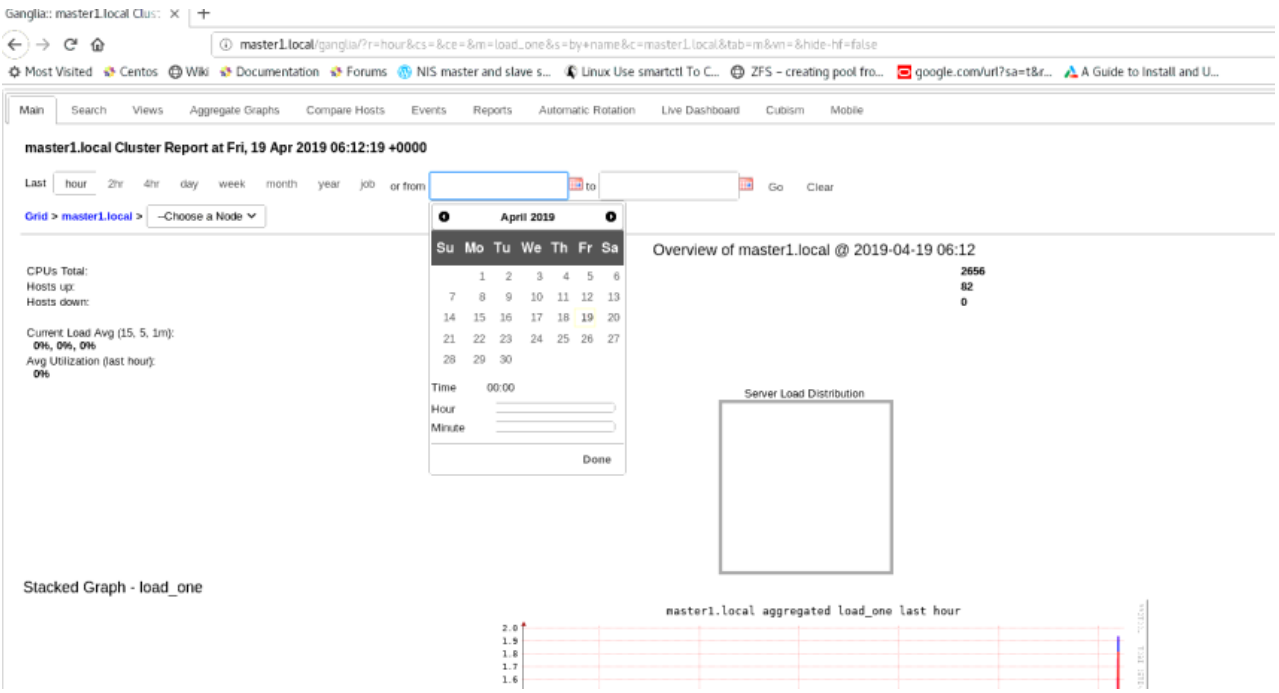
Uplink Connection

To Monitor the Nodes, open the web-browser and type in the following URL,
http://master1.local/ganglia

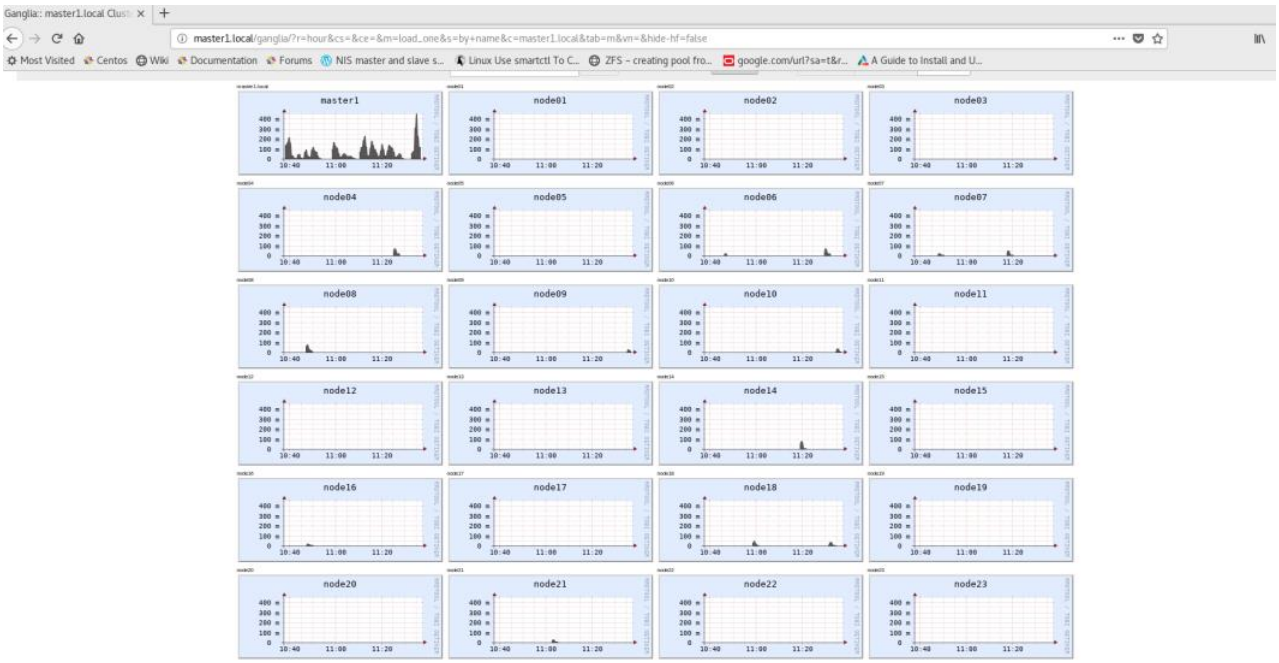
Hover the curser on a particular feature and click on the feature to get more information on that feature, The below figure shows the details shown after selecting Team centre feature.



To filter the information for a range of date select from and to date on the top of the page as shown in the figure below.



Node loads



What is PBS Professional?

PBS Professional software optimizes job scheduling and workload management in high-performance computing (HPC) environments – clusters, clouds, and supercomputers –

improving system efficiency and people's productivity. Built by HPC people for HPC people, PBS Professional is fast, scalable, secure, and resilient, and supports all modern infrastructure, middleware, and applications.

The **pbsnodes** command is used to mark nodes down, free or off line. It can also be used to list nodes and their state. Node information is obtained by sending a request to the PBS job server.

PBS Node Details (The attributes include "state" and "properities")

```
[root@master1 ~]# pbsnodes -a
node01
Mom = node01
Port = 15002
pbs_version = 19.0.0
ntype = PBS
state = free
pcpus = 32
resources_available.arch = linux
resources_available.host = node01
resources_available.mem = 97421020kb
resources_available.ncpus = 32
resources_available.vnode = node01
resources_assigned.accelerator_memory = 0kb
resources_assigned.hbmem = 0kb
resources_assigned.mem = 0kb
resources_assigned.naccelerators = 0
resources_assigned.ncpus = 0
resources_assigned.vmem = 0kb
resv_enable = True
sharing = default_shared
last_state_change_time = Wed Apr 10 18:42:23 2019
```

Submitting Job

```
[locuz@master1 32node]$ qsub job.pbs
```

1213.master

Check the status of job

[locuz@master1 gcc]$ qstat

```
Job id          Name            User            Time Use S Queue
---------------  ---------------- ----------------  -------- - -----
1197.master      Test            locuz                  0 Q workq      (The Job is in queue)
```

[locuz@master1 32node]$ qstat

```
Job id          Name            User            Time Use S Queue
---------------  ---------------- ----------------  -------- - -----
1213.master      lammps-8nodes   locuz           00:00:00 R workq ( The Job is running )
```

[root@master2 ~]# pbsnodes –avS (list the node status)

```
vnode        state        OS     hardware host        queue       mem    ncpus  nmics  ngpus  comment
---------------  ---------------  --------  --------  ----------------  ----------  --------  -------  -------  -------  ---------
node01       free         --     --     node01       --          93gb   32     0      0 --
node02       free         --     --     node02       --          93gb   32     0      0 --
node03       free         --     --     node03       --          93gb   32     0      0 --
node04       free         --     --     node04       --          93gb   32     0      0 --
node05       free         --     --     node05       --          93gb   32     0      0 --
node06       free         --     --     node06       --          93gb   32     0      0 --
node07       free         --     --     node07       --          93gb   32     0      0 --
node08       free         --     --     node08       --          86gb   32     0      0 --
node09       free         --     --     node09       --          93gb   32     0      0 --
node10       free         --     --     node10       --          93gb   32     0      0 --
.......
```

Delete the job

[locuz@master1 gcc]$ qdel 1197 (1197 is Job ID)

Verify PBS

- On the master node

/etc/init.d/pbs status

/etc/init.d/pbs start (to start manually)

- On the compute node

/etc/init.d/pbs status

etc/init.d/pbs start (to start manually)

Shutdown procedure for the cluster:

- Plan the Shutdown activity and inform the Cluster Users of the Schedule for the shutdown.
- Verify if any jobs are running and kill all the running jobs.
- Power off all the compute nodes.
  # pdsh "poweroff"
- Power off the Master node
  #poweroff

ZFS (Z File System)

ZFS is a local file system and logical volume manager created by Sun Microsystems Inc. to direct and control the placement, storage and retrieval of data in enterprise-class computing systems.

ZFS administration

lssci (list SCSI disks)

zpool status (list zpool status)

zpool list (Zpool list)

zpool export master-arc (Exporting pool)

zpool import -f master-arc (Import pool to master)

zpool import archive-storage (run it in Compute storage)

zfs mount –a (ZFS mount)

ZFS Quota

[root@storage ~]# zfs userspace home-storage/home

```
TYPE       NAME         USED  QUOTA  OBJUSED  OBJQUOTA
POSIX User  bhanupb     58.8K    5T       8      none
POSIX User  bhupendrard 58.8K    5T       8      none
POSIX User  conserver    135K  none      14      none
POSIX User  dheerajp     58.8K    5T       8      none
POSIX User  indrajittah  58.8K    5T       8      none
POSIX User  jmondal      58.8K    5T       8      none
POSIX User  kallolp      58.8K    5T       8      none
POSIX User  locuz         628G    5T   23.4K      none
POSIX User  locuztest     148K    5T      21      none
POSIX User  mrinmoym     58.8K    5T       8      none
POSIX User  navdeepr     89.8K    5T      15      none
POSIX User  ndube        58.8K    5T       8      none
POSIX User  perlekar     58.8K    5T       8      none
POSIX User  pghosh       90.3K    5T      16      none
POSIX User  pnath        58.8K    5T       8      none
POSIX User  rajsekhard   58.8K    5T       8      none
POSIX User  rashmir      58.8K    5T       8      none
POSIX User  rayanc       58.8K    5T       8      none
POSIX User  root         19.0G  none      45      none
POSIX User  smarajit     58.8K    5T       8      none
POSIX User  surajit      58.8K    5T       8      none
POSIX User  tapass       58.8K    5T       8      none
POSIX User  test1        80.6K    5G      15      none
POSIX User  test2        58.8K    1G       8      none
POSIX User  test3        1.00G    1G      16      none
POSIX User  viakshp      58.8K    5T       8      none
POSIX User  vishnuvk     58.8K    5T       8      none
```

Basic Cluster Trouble shooting

NIS:

- Check if nis services are running on the master node

#/etc/init.d/ypserv status

#/etc/init.d/ypserv start to start manually if it is stopped

- Check if nis service is running on the compute node

#/etc/init.d/ypbind status

#/etc/init.d/ypbind start to start manually.

For User addition please use below command

```
[root@master1 ~]# which adduser_nis
/usr/bin/adduser_nis
```

```
[root@master1 ~]# adduser_nis
Please enter the username you want to add.
test1111
Changing password for user test1111.
New password:
BAD PASSWORD: The password contains the user name in some form
Retype new password:
passwd: all authentication tokens updated successfully.
gmake[1]: Entering directory `/var/yp/local'
gmake[1]: `ypservers' is up to date.
gmake[1]: Leaving directory `/var/yp/local'
gmake[1]: Entering directory `/var/yp/local'
Updating passwd.byname...
master2.local: Master's version not newer
master1.local: Master's version not newer
Updating passwd.byuid...
master2.local: Master's version not newer
master1.local: Master's version not newer
Updating group.byname...
master2.local: Master's version not newer
master1.local: Master's version not newer
Updating group.bygid...
master2.local: Master's version not newer
master1.local: Master's version not newer
Updating netid.byname...
master2.local: Master's version not newer
master1.local: Master's version not newer
Updating shadow.byname...
master2.local: Master's version not newer
master1.local: Master's version not newer
gmake[1]: Leaving directory `/var/yp/local'
 Quota Enabled 5 TB
POSIX User  test1111     58.8K     5T        8       none
[root@master1 ~]#
```

Infiniband

check the ibv_devinfo command and verify the status of the infiniband.

#ibv_devinfo

verify the power to the Infiniband switch and the status of the LED on the IB Switch.

Home directory

- Verify is the /home directory is mounted from the lustre filesystem

#pdsh –a “df –Th”

- If the filesystem is not mounted mount it manually using the following command.

#modprobe lustre

#mount –t nfs 192.168.12.3:/home-storage/home /home

- verify the status of openibd

#/etc/init.d/openibd status

#/etc/init.d/openibd start

verify the license server

- Check if the license server is communicating to the nodes by using the ping command from the compute nodes to the license server.

Power On the cluster:

- Power on the Ethernet and the Infiniband Switches and wait for 15 minutes
- Power on the master node and wait until it boots up.
- After the master node is completely booted, and wait for 15 minutes.
- Power on the compute nodes one by one.

Access the nodes via IPMI for remote management

Open the web-browser and enter the IPMI IP (http://192.168.102.x) of the node to manage the node remotely.

Recommendation

- Please maintain required cooling (< 20 degree Celsius) round the clock inside DC.
- Please maintain grounding for each rack less than 1v round the clock.
- Please provide uninterrupted power supply round the clock for all equipment installed in DC.
- Please do not change configuration done by us.
- Please avoid logging from root.
- For any assistance please connect Locuz Help desk toll free (1-800-103-6971) or raise an service request at servicedesk@locuz.com .