ACMAD HPC

System Administrator manual

December 01, 2021
Author: eXact lab

Abstract

This document describes the installation procedures and configuration activities for the High Performance Computing facility installed at ACMAD in June 2021. A brief description of hardware is given; then services installed on the system are explained from a system administration point of view. The document is meant as an operations manual for the System Administrator personnel in charge of maintaining the infrastructure and associated services.

Table of contents

Hardware inventory

The environment is based on DELL hardware and is composed of 13 servers, 4 Storage Area Network, a Gigabit Ethernet switch, and an InfiniBand switch.

Servers for computing and services

- 2 x master DELL R740
- 2 x login DELL R740
- 4 x lustre DELL R740
- 4 x compute DELL R640
- 1 x monitoring DELL R640

Please find the quick specs of these servers here:

- R740: https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-r740-spec-sheet.pdf
- R640: https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-r640-spec-sheet.pdf

While the technical specs are as follows:

Server	**CPU**	**RAM**	**Network**	**Local Storage** **SAS cards to SAN**
2 x Master	2 x Intel Xeon Gold 6134 3.2G (2 x 8 core)	192 GB	2 x 10Gb 2 x 1Gb 2 x InfiniBand	2 x 2TB HDD 2 x SAS (2 x 12Gbps)
2 x Login	2 x Intel Xeon Gold 6134 3.2G (2 x 8 core)	96 GB	2 x 10Gb 2 x 1Gb 2 x InfiniBand	2 x 2TB HDD
4 x Lustre	2 x Intel Xeon Gold 6136 3.2G (2 x 12 core)	384 GB	4 x 1Gb 1 x InfiniBand	2 x 2TB HDD 2 x SAS (2 x 12Gbps)
4 x Compute	2 x Intel Xeon Gold 6248 2.5G	192 GB	4 x 1Gb 2 x InfiniBand	1 x 480GB SSD

	(2 x 20 core)			
1 x Monitoring	2 x Intel Xeon Gold 5118 2.3G (2 x 12 core)	96 GB	4 x 1Gb 1 x InfiniBand	2 x 480 GB SSD
13 servers	**344 cores**	**3 TB**	**-**	**35 TB local storage** **694 TB on SAN**

Please note that:

- InfiniBand is a MT 4119 ConnectX-5 EDR adapter - 100 Gbit/second
- All the servers have hypertreading enabled

Storage SAN

- 1 x DELL EMC PowerVault 4012
- 1 x DELL EMC PowerVault 4024
- 2 x DELL EMC PowerVault 4084

Please find the quick specs of the DELL PowerVault ME4 series here:
https://www.dell.com/en-en/work/shop/productdetailstxn/powervault-me4-series

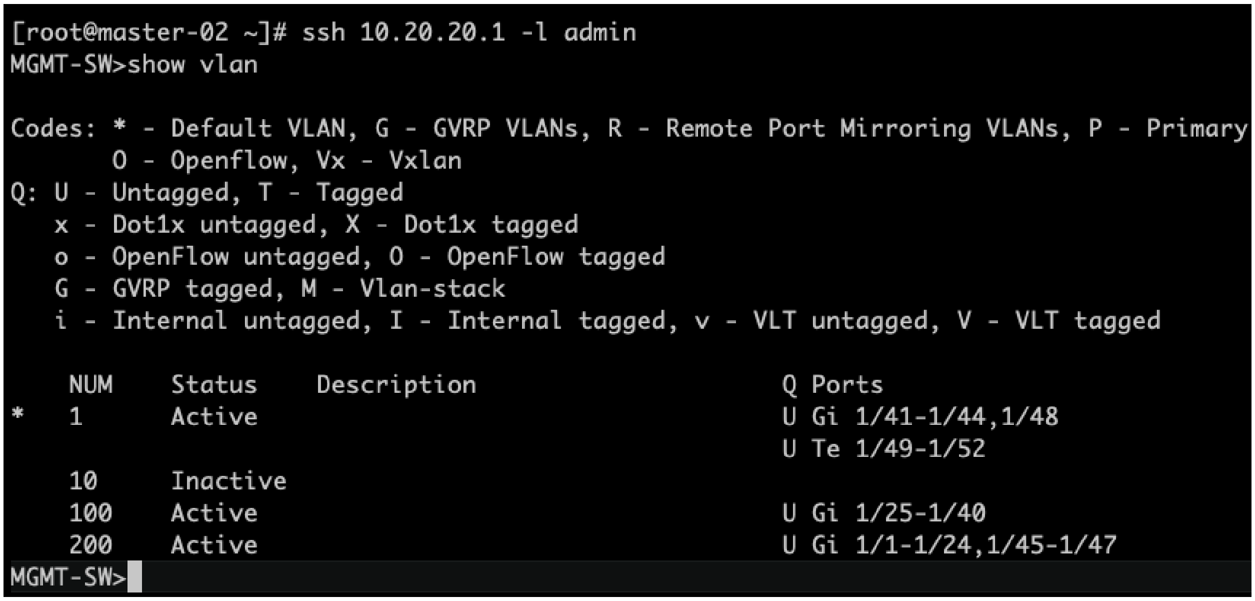
Networking

- 1 x DELL S3048 Gigabit Ethernet Switch: spec sheet here.
- 1 x Mellanox SB7800 InfiniBand Switch: spec sheet here.

Network design

The network layer has been segmented in different VLANs to accomplish different services.
The DELL S3048 has been divided into the following VLANs:

- **1 - default**, used for external connectivity and uplinking
- **100 - deployment** / in-band management, used for deployment and provisioning of operating systems on servers
- **200 - out-of-band** management, used for BMCs, switches and SAN controllers management

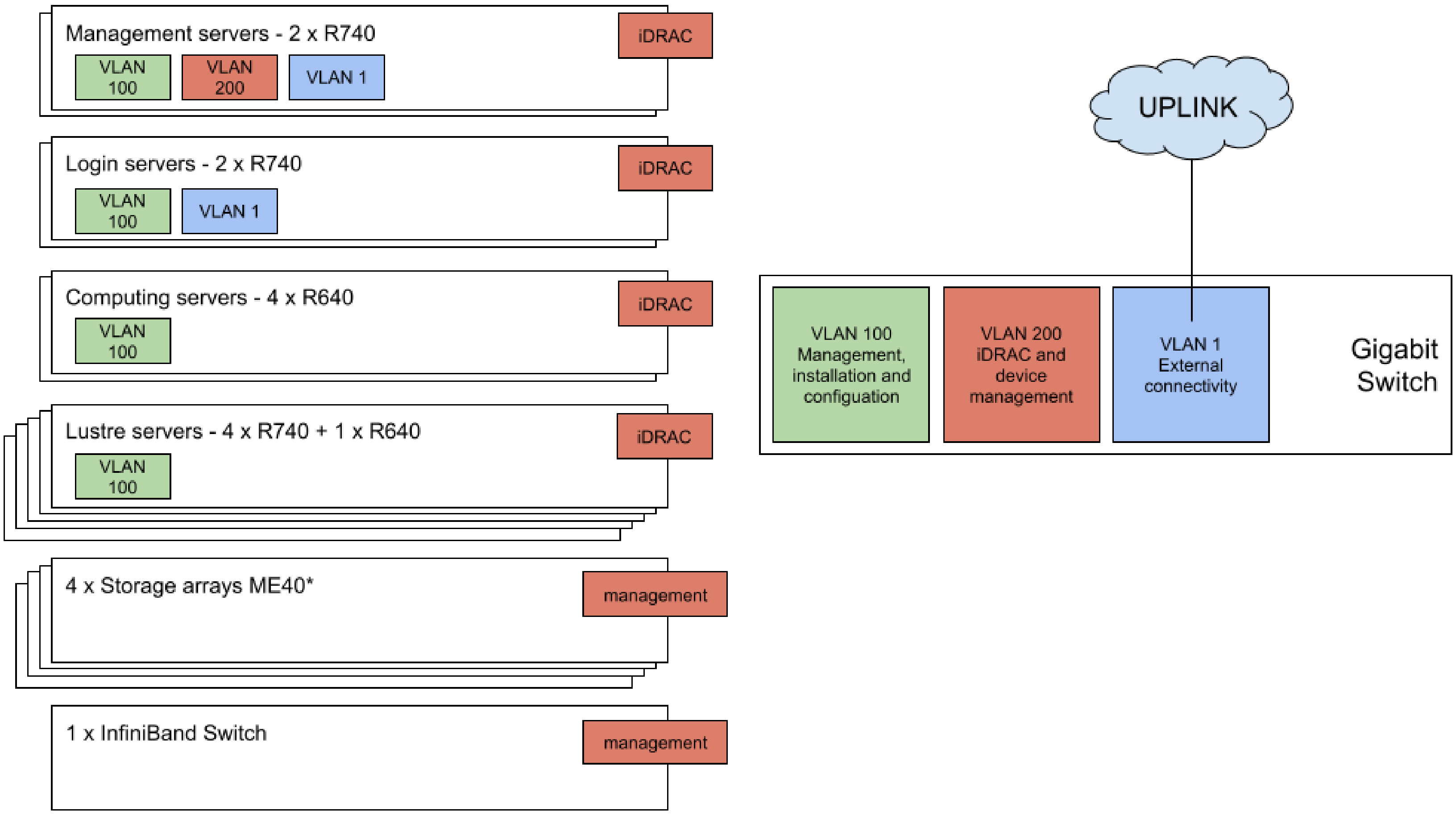


```
[root@master-02 ~]# ssh 10.20.20.1 -l admin
MGMT-SW>show vlan

Codes: * - Default VLAN, G - GVRP VLANs, R - Remote Port Mirroring VLANs, P - Primary
       O - Openflow, Vx - Vxlan
Q: U - Untagged, T - Tagged
   x - Dot1x untagged, X - Dot1x tagged
   o - OpenFlow untagged, O - OpenFlow tagged
   G - GVRP tagged, M - Vlan-stack
   i - Internal untagged, I - Internal tagged, v - VLT untagged, V - VLT tagged

    NUM    Status    Description                        Q Ports
*   1      Active                                       U Gi 1/41-1/44,1/48
                                                        U Te 1/49-1/52
    10     Inactive
    100    Active                                       U Gi 1/25-1/40
    200    Active                                       U Gi 1/1-1/24,1/45-1/47
MGMT-SW>
```

Each component of the HPC facility has been connected to one or more VLAN. The high level design of the cluster's network infrastructure is shown in the following picture.


Management servers - 2 x R740
VLAN 100
VLAN 200
VLAN 1
iDRAC
Login servers - 2 x R740
VLAN 100
VLAN 1
iDRAC
Computing servers - 4 x R640
VLAN 100
iDRAC
Lustre servers - 4 x R740 + 1 x R640
VLAN 100
iDRAC
4 x Storage arrays ME40*
management
1 x InfiniBand Switch
management
UPLINK
VLAN 100 Management, installation and configuation
VLAN 200 iDRAC and device management
VLAN 1 External connectivity
Gigabit Switch

HLD of cluster network

To know exactly where each server or device has been connected, please check the following table.

Port number S3048	Port type	Host	Port/NIC	MAC address	Vlan
1	1Gb	me4084-1-a	-	00:C0:FF:52:18:F9	200
2	1Gb	me4084-1-b	-	00:c0:ff:51:e2:da	200
3	1Gb	me4084-2-a	-	00:c0:ff:51:e4:13	200
4	1Gb	me4084-2-b	-	00:c0:ff:51:df:1f	200
5	1Gb	oss-02	iDRAC	2c:ea:7f:55:bf:50	200
6	1Gb	oss-01	iDRAC	2c:ea:7f:56:08:b9	200
7	1Gb	mds-02	iDRAC	2c:ea:7f:55:cf:ab	200
8	1Gb	mds-01	iDRAC	2c:ea:7f:55:d7:94	200
9	1Gb	iml	iDRAC	34:48:ed:f1:80:5c	200
10	1Gb	me4024-1-b	-	00:c0:ff:51:e2:5e	200
11	1Gb	me4024-1-a	-	00:c0:ff:51:d2:50	200
12	1Gb	login-01	iDRAC	70:b5:e8:cd:fb:78	200
13	1Gb	login-02	iDRAC	70:b5:e8:cd:db:b0	200
14	1Gb	master-01	iDRAC	70:b5:e8:cd:db:68	200
15	1Gb	master-02	iDRAC	70:b5:e8:cd:db:1a	200
16	1Gb	compute-01	iDRAC	34:48:ed:eb:48:12	200
17	1Gb	compute-02	iDRAC	34:48:ed:eb:4c:f2	200
18	1Gb	compute-03	iDRAC	34:48:ed:eb:46:da	200
19	1Gb	compute-04	iDRAC	34:48:ed:eb:46:fe	200
20	1Gb	me4012-1-b	-	00:c0:ff:51:d0:fe	200
21	1Gb	me4012-1-a	-	00:c0:ff:51:df:a9	200
22	1Gb	master-01	eno4	bc:97:e1:5a:c6:5f	200
23	1Gb	master-02	eno4	BC:97:E1:5A:C8:0F	200
24	1Gb			0c:42:a1:e5:c0:64	200
25	1Gb	monitoring / iml	em1	F0:D4:E2:EB:FB:98	100
26	1Gb	mds-01	em1	34:48:ed:f3:d7:54	100
27	1Gb	mds-02	em1	34:48:ed:f3:d1:fc	100
28	1Gb	oss-01	em1	34:48:ed:f3:de:98	100
29	1Gb	oss-02	em1	34:48:ed:f3:e1:b8	100
30	1Gb				100
31	1Gb	login1	em3	bc:97:e1:5a:d2:64	100
32	1Gb	login2	em3	bc:97:e1:5a:ae:d0	100
33	1Gb	master-01	eno3	bc:97:e1:5a:c6:5e	100
34	1Gb	master-02	eno3	bc:97:e1:5a:c8:0e	100

35	1Gb				100
36	1Gb				100
37	1Gb	compute-01	em1	f0:d4:e2:eb:b7:cc	100
38	1Gb	compute-02	em1	f0:d4:e2:eb:b2:f4	100
39	1Gb	compute-03	em1	f0:d4:e2:eb:ca:50	100
40	1Gb	compute-04	em1	F0:D4:E2:EC:00:18	100
41	1Gb				1
42	1Gb				1
43	1Gb				1
44	1Gb				1
45	1Gb				200
46	1Gb				200
47	1Gb	monitoring / iml	eno2	f0:d4:e2:eb:fb:99	200
48	1Gb	UPLINK			1
49	10Gb	login1	em1	bc:97:e1:5a:d2:66	1
50	10Gb	login2	em1	bc:97:e1:5a:ae:d2	1
51	10Gb	master1	eno1np0	bc:97:e1:5a:c6:60	1
52	10Gb	master2	eno1np0	bc:97:e1:5a:c8:10	1

A different IP address space has been chosen for the different VLANs.

- **1 - default**. To be used for external connectivity. **192.168.0.0/24**
- **100 - deployment**. To be used for in-band management. **10.10.0.0/24**
- **200 - management**. To be used for out-of-band management. **10.20.0.0/24**

Please check the following table to understand how these address spaces have been used to connect the HPC cluster components on the networks.

Hostname	**VLAN 100**	**VLAN 200**	**VLAN 1**	**InfiniBand**
master-01	10.10.0.1 10.10.10.1 VIP	10.20.0.1 10.20.0.101	192.168.0.81 192.168.0.80 VIP	10.60.0.1 10.60.10.1
master-02	10.10.0.2	10.20.0.2 10.20.0.102	192.168.0.82	10.60.0.2
login-01	10.10.0.3 10.10.10.3 VIP	10.20.0.3	192.168.0.83 192.168.0.85 VIP	10.60.0.3
login-02	10.10.0.4	10.20.0.4	192.168.0.84	10.60.0.4
oss-01	10.10.0.5	10.20.0.5		10.60.0.5

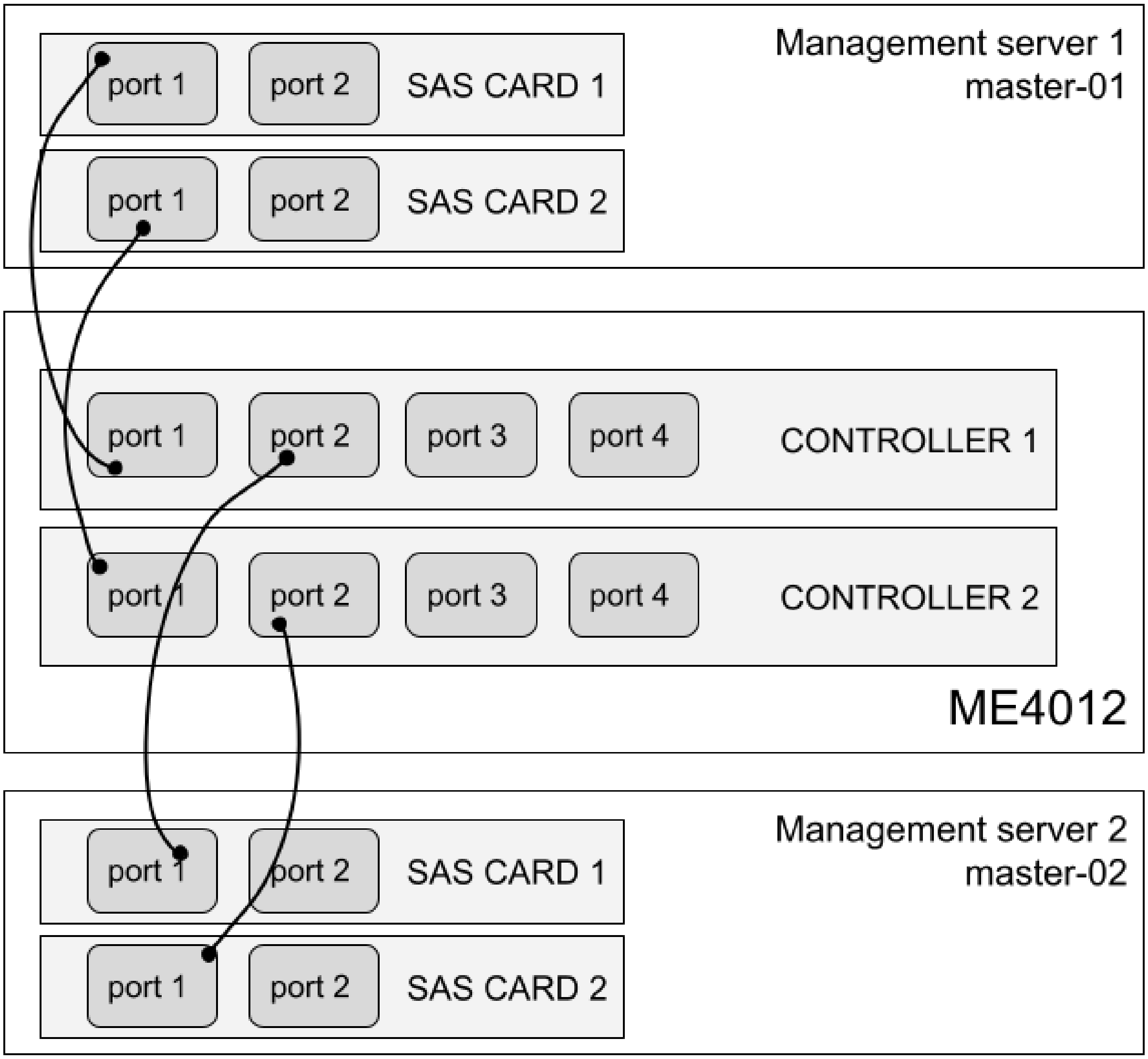
oss-02	10.10.0.6	10.20.0.6		10.60.0.6
mds-01	10.10.0.7	10.20.0.7		10.60.0.7
mds-02	10.10.0.8	10.20.0.8		10.60.0.8
monitoring / iml	10.10.0.9	10.20.0.9		10.60.0.9
compute-01	10.10.0.10	10.20.0.10		10.60.0.10
compute-02	10.10.0.11	10.20.0.11		10.60.0.11
compute-03	10.10.0.12	10.20.0.12		10.60.0.12
compute-04	10.10.0.13	10.20.0.13		10.60.0.13
MGNT-SW		10.20.20.1		
mnlx-sw		10.20.20.2		
me4012-1-b		10.20.10.11		
me4024-1-b		10.20.10.22		
me4084-1-b		10.20.10.33		
me4084-2-b		10.20.10.44		
me4012-1-a		10.20.10.1		
me4024-1-a		10.20.10.2		
me4084-1-a		10.20.10.3		
me4084-2-a		10.20.10.4		

Storage

The ACMAD HPC has four SAN based on Dell PowerVault technologies, for a total of 694 TB RAW. Common specs of all the storage systems are described here.
Please note on every SAN the firmware has been upgraded from version *GT280R006-02* to version *GT280R008-04* (released November 2020).

SAN	**Disks**	**Raidpool**	**Connected to**
ME4012	6 x 2.4TB HDD	1 x RAID6	master-01, master-02
ME4024	12 x 960GB SSD	1 x RAID10 2 hot spare disks	mds-01, mds-02
ME4084	84 x 4TB HDD	8 x RAID6 (10 disks each) 4 hot spare disks	oss-01, oss-02
ME4084	84 x 4TB HDD	8 x RAID6 (10 disks each) 4 hot spare disks	oss-01, oss-02

Each SAN is connected to 2 servers. The following picture describes how the master servers have been connected to the ME4012.



The configuration is the same on the ME4024, connected to the Lustre mds servers.

The 2 Lustre oss servers are connected to 2 x ME4084. On these servers both the ports of each SAS card have been connected, in order to have redundancy at server level, SAS cards, SAN controllers.

The LUN presentation of all the storage in the ACMAD HPC cluster is described in the following table. The multipath configuration is the same on all the servers and written in /etc/multipath.conf.

SAN	Connected to hosts	Volumes / LUN	Multipath names
ME4012	Master servers	2 x 2TB	/dev/mapper/home /dev/mapper/cm-shared
ME4024	Lustre mds servers	1 x 4.4TB 1 x 250MB	/dev/mapper/mdt /dev/mapper/mgt
ME4084 - 1	Lustre oss servers	8 x 29TB	/dev/mapper/ost01 /dev/mapper/ost02 ... /dev/mapper/ost09
ME4084 - 2	Lustre oss servers	8 x 29TB	/dev/mapper/ost10 /dev/mapper/ost11 ... /dev/mapper/ost16

Path redundancy - multipath device mapper

The multipath service allows the combination of multiple physical connections in one single virtual device. On each disk array, every LUN is exposed by both the controllers to each server to which is connected. This means that every LUN is seen twice on each server (e.g. LUN 1 is seen on /dev/sdc and /dev/sdd). The multipath service aggregates both the paths in a single virtual device, whose name is /dev/mapper/[name of device].

ACMAD HPC Cluster configuration

All the servers of ACMAD HPC Cluster have been installed with RedHat 7.9, and configured to expose different services. These services are briefly described as follows, and detailed in the following sections.

- 2 x master servers: allow the system administrators to control all the cluster services. They aren't meant to be used by normal users, and privileged access is needed. They store all the configurations and installation procedures needed to re-install or re-configure the cluster.
- 2 x login servers: allow the scientific users to log in on the Cluster, by means of SSH, and submit computational jobs on the Slurm queue.
- 4 x compute servers: allow the computational jobs to run on high performance CPUs.
- 4 x Lustre servers: allow the disk arrays to expose all their disks under a common, redundant, high performance namespace.
- 1 x monitoring server: allow monitoring of all the infrastructure.

Master servers

The master servers have been installed with an ISO image of Bright Cluster Manager , version 9.1, based on RedHat 7.9. These servers have been configured in High Availability, managed by Bright CM itself. To acquire further knowledge of the Bright product, please read official documentation available here.

The master servers provide the following services, managed by Bright CM, and needed to maintain the cluster in full efficiency:

- DHCP: used to assign IP address on all the cluster networks
- DNS: used for name resolution
- PXE / Netboot: used for operating system provisioning on all the nodes
- Bright portal: GUI to manage and monitor all the nodes of the cluster
- cmsh: Bright Cluster Manager shell, used as an alternative to the GUI to manage all the above services
- Slurm server: queue system to allow scheduling of computational jobs on the computing nodes
- Gateway: all the servers in the HPC cluster use the master servers as gateways to route traffic on external networks

Networks on master servers

The master servers manage and have access to all the networks of the HPC cluster. They have:

- First 10 Gigabit interface configured on 192.168.0.0/16 network (VLAN 1), to allow system administrators access from external
- Second 10 Gigabit interface is unplugged
- First 1 Gigabit interface configured on 10.10.0.0/16 network (VLAN 100), to allow in-band management
- Second 1 Gigabit interface configured on 10.20.0.0/16 network (VLAN 200), to allow out-of-band management
- InfiniBand interface configured on 10.60.0.0/16

Out-of-band management

Out-of-band management is granted after access on the master servers. Once logged in on the master servers, the system administrator can control all the devices in the HPC Cluster:

- All the iDRAC of all the servers
- All the management interface of all the disk arrays (each disk array has two controllers, hence two management interfaces)
- The management of the Ethernet and InfiniBand switches

The IP address of all the cluster components is shown in the following table. Please note that as a security measure, password has not been written on the documentation.
On each iDRAC, passwordless access is granted from the master servers, since a public key has been deployed on all the iDRAC.

Hostname	Out-of-band (iDRAC or management interface)	Username	Password
master-01	10.20.0.1	root	Please ask for password!
master-02	10.20.0.2		
login-01	10.20.0.3		
login-02	10.20.0.4		
oss-01	10.20.0.5		
oss-02	10.20.0.6		
mds-01	10.20.0.7		
mds-02	10.20.0.8		
iml	10.20.0.9		
compute-01	10.20.0.10		
compute-02	10.20.0.11		
compute-03	10.20.0.12		
compute-04	10.20.0.13		
MGNT-SW	10.20.20.1	admin	
mnlx-sw	10.20.20.2		
me4012-1-b	10.20.10.11	administrator	
me4024-1-b	10.20.10.22		
me4084-1-b	10.20.10.33		
me4084-2-b	10.20.10.44		
me4012-1-a	10.20.10.1		
me4024-1-a	10.20.10.2		
me4084-1-a	10.20.10.3		
me4084-2-a	10.20.10.4		

Failover groups

A failover group is a set of two servers in an active/standby configuration. The services configured in a failover group run exclusively on the active server, and the users connect to the active server to consume them. In case of failure of the active, the standby server will take over the services, ensuring continuity for the users.

The master servers and the login servers have been configured in different failover groups:

- The failover group on the master servers allow different vital services to be still alive in case of failure of one of the two servers. These services are: DHCP, DNS, PXE/Netboot, users' home, scientific software, Slurm.
- The failover group on the login servers allow login of the scientific users to be still available in case of failure of one of the login servers.

In order to guarantee continuity of service, each failover group must have a floating IP address belonging to the active server. In case of failure, the floating IP will move to the standby server, which will be promoted as the active one.

The following table displays how the failover groups have been used to create a single point of access on each network.

Failover group	**IP on primary**	**IP on secondary**	**Floating IP (VIP)**
master - external access	192.168.0.81	192.168.0.82	192.168.0.80
master - in-band network	10.10.0.1	10.10.0.2	10.10.10.1
master - infiniband network	10.60.0.1	10.60.0.2	10.60.10.1
login - external access	192.168.0.83	192.168.0.84	192.168.0.85

Users must adopt the Floating IP for their SSH connection. Primary or secondary IP must be used only in case of diagnostic actions and/or issues.

Please note that the Floating IP of the master servers on the in-band and InfiniBand network are also gateways for all the other servers in the HPC cluster. This means that the login servers and the compute node will use 10.10.10.1 or 10.60.10.1 as gateway for the in-band and InfiniBand networks.

Storage configuration

The two master servers have 2 x 2TB local disks each, configured in a RAID1 volume. On this volume the operating system has been installed.

The master servers have been connected to the ME4012 disk array, where two different volumes have been created for:

- Users’ home: where the users have their home, once logged in on the login nodes. The home of the users are then exported by means of NFS to the login and computing nodes. This volume is 2TB big, but can be increased in size, as needed.
- Bright Cluster Manager: the Bright CM uses some shared storage to store configurations and other important files. This volume is also used to store scientific software and then exported via NFS to all the computing nodes. This volume is 2TB big, but can be increased in size, as needed.

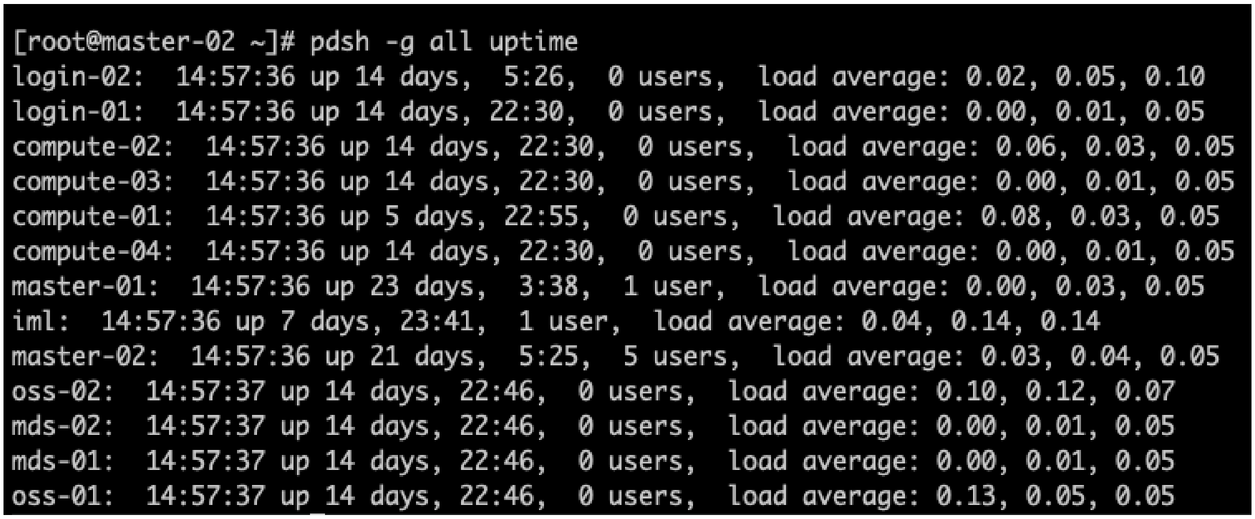
How to access the master servers

From the external network, access on the active master server is granted on IP 192.168.0.80 by means of SSH. We recommend granting access only to experienced system administrators. We recommend avoiding password access, and to configure a passwordless environment to avoid security issues.
The Bright Cluster Manager GUI is available on the same IP address, by means of connection to http://192.168.0.80.

Main operations on the master servers

Parallel shell

The parallel shell `pdsh` (https://github.com/chaos/pdsh) is installed by Bright CM, and available for the system administrator. Please check `/etc/genders` for the pdsh groups available.



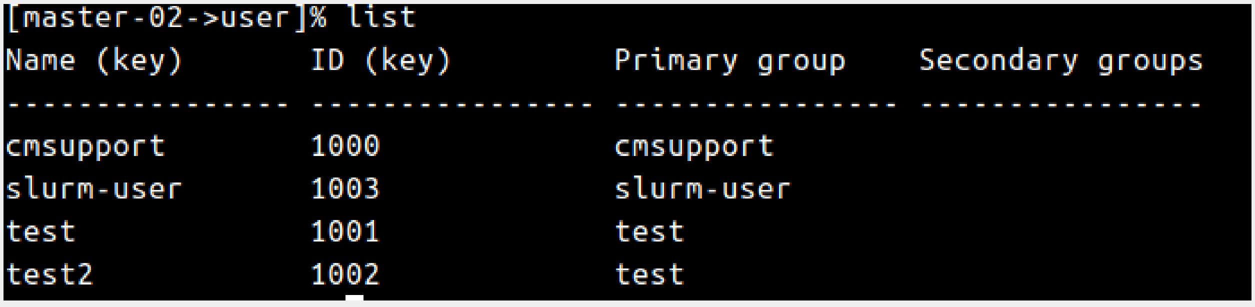
```
[root@master-02 ~]# pdsh -g all uptime
login-02:  14:57:36 up 14 days,  5:26,  0 users,  load average: 0.02, 0.05, 0.10
login-01:  14:57:36 up 14 days, 22:30,  0 users,  load average: 0.00, 0.01, 0.05
compute-02:  14:57:36 up 14 days, 22:30,  0 users,  load average: 0.06, 0.03, 0.05
compute-03:  14:57:36 up 14 days, 22:30,  0 users,  load average: 0.00, 0.01, 0.05
compute-01:  14:57:36 up 5 days, 22:55,  0 users,  load average: 0.08, 0.03, 0.05
compute-04:  14:57:36 up 14 days, 22:30,  0 users,  load average: 0.00, 0.01, 0.05
master-01:  14:57:36 up 23 days,  3:38,  1 user,  load average: 0.00, 0.03, 0.05
iml:  14:57:36 up 7 days, 23:41,  1 user,  load average: 0.04, 0.14, 0.14
master-02:  14:57:36 up 21 days,  5:25,  5 users,  load average: 0.03, 0.04, 0.05
oss-02:  14:57:37 up 14 days, 22:46,  0 users,  load average: 0.10, 0.12, 0.07
mds-02:  14:57:37 up 14 days, 22:46,  0 users,  load average: 0.00, 0.01, 0.05
mds-01:  14:57:37 up 14 days, 22:46,  0 users,  load average: 0.00, 0.01, 0.05
oss-01:  14:57:37 up 14 days, 22:46,  0 users,  load average: 0.13, 0.05, 0.05
```

Node provisioning

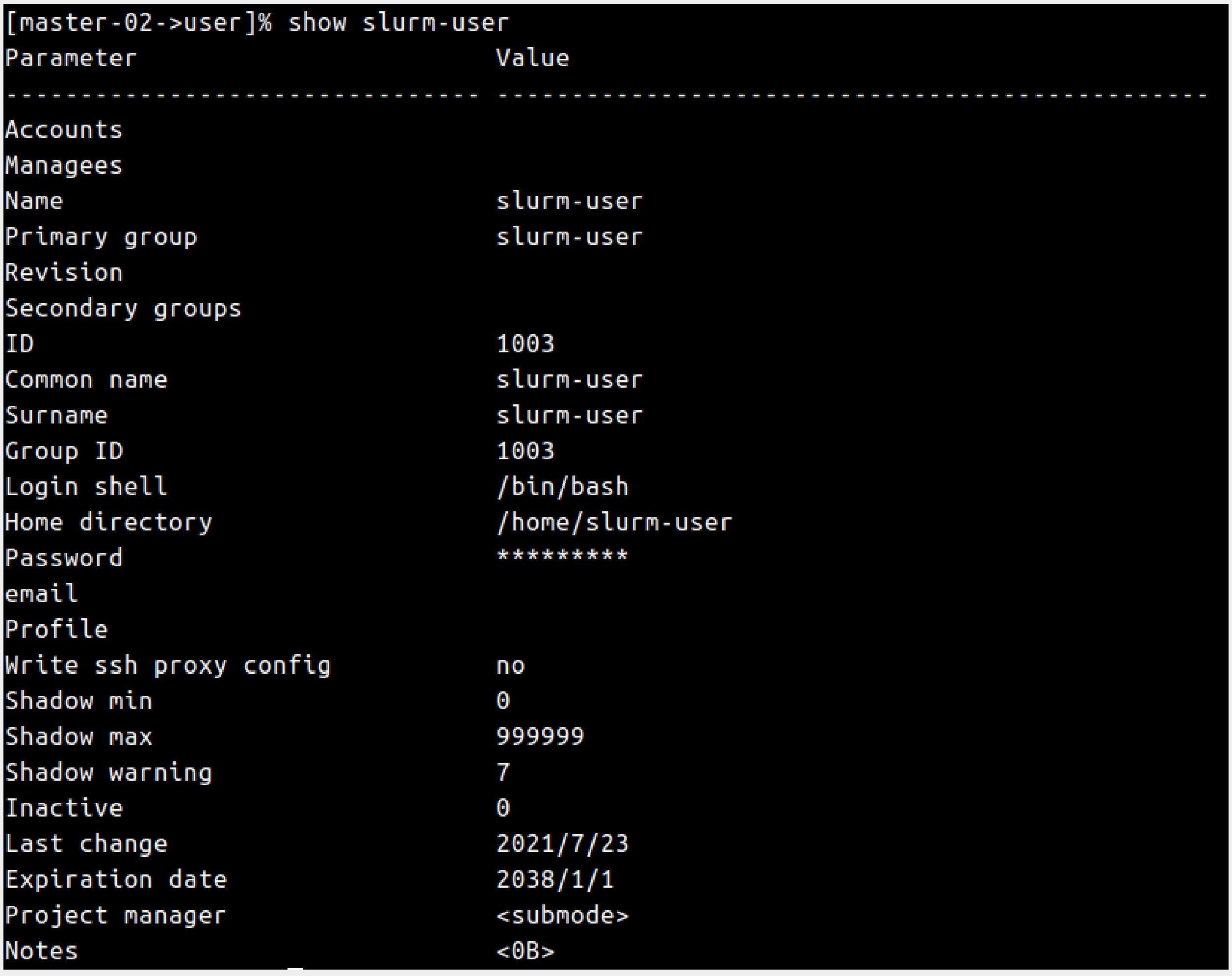
Bright

How to create users from master node

```
#cmsh
#users
#add test-user
#show test-user
## set user parameters (Password; Home directory ecc):
#set test-user password [password]
#commit
```



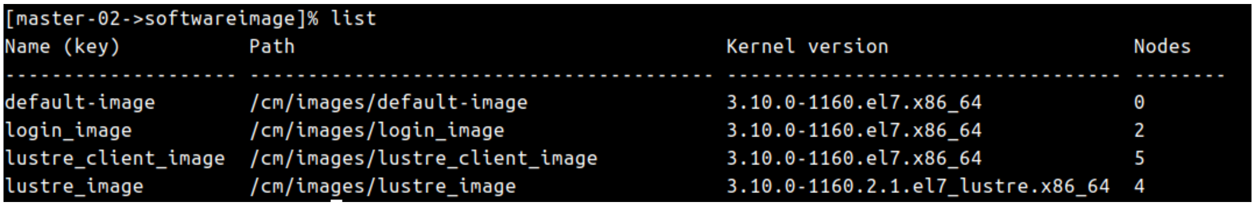
```
[master-02->user]% list
Name (key)       ID (key)         Primary group    Secondary groups
---------------- ---------------- ---------------- ----------------
cmsupport        1000             cmsupport
slurm-user       1003             slurm-user
test             1001             test
test2            1002             test
```



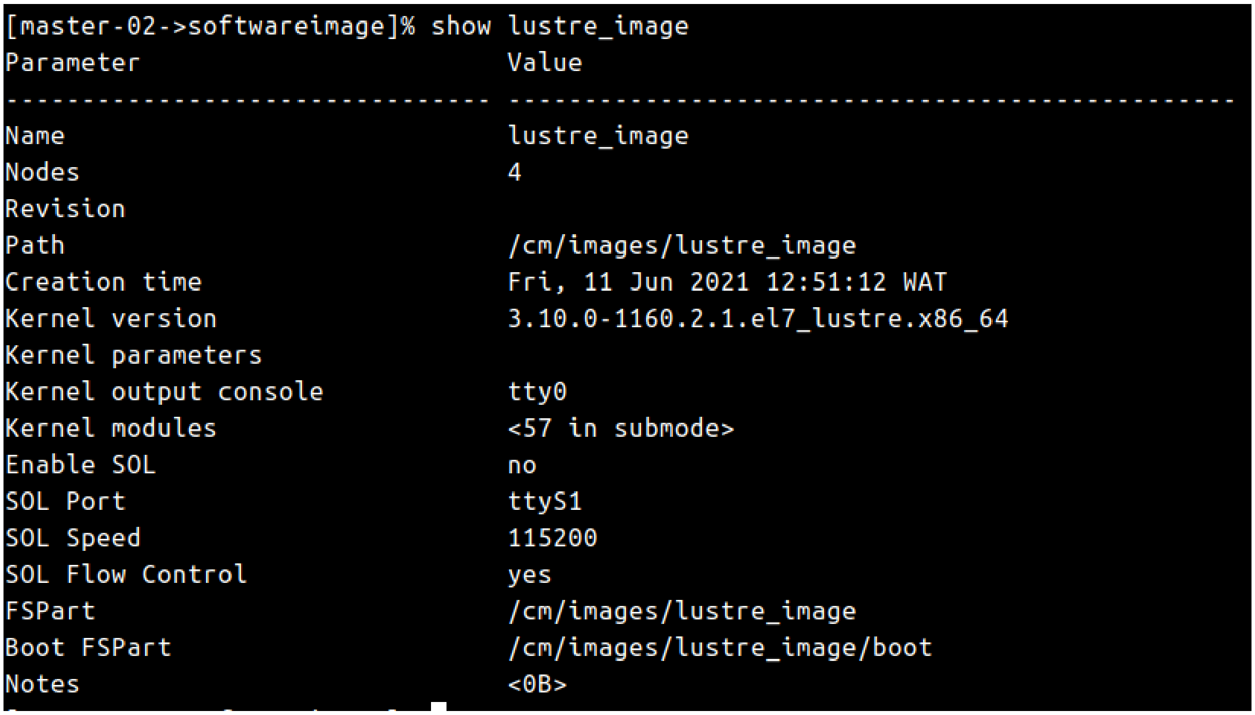
```
[master-02->user]% show slurm-user
Parameter                        Value
-------------------------------- ------------------------------------------------
Accounts
Managees
Name                             slurm-user
Primary group                    slurm-user
Revision
Secondary groups
ID                               1003
Common name                      slurm-user
Surname                          slurm-user
Group ID                         1003
Login shell                      /bin/bash
Home directory                   /home/slurm-user
Password                         *********
email
Profile
Write ssh proxy config           no
Shadow min                       0
Shadow max                       999999
Shadow warning                   7
Inactive                         0
Last change                      2021/7/23
Expiration date                  2038/1/1
Project manager                  <submode>
Notes                            <0B>
```

The cluster manager used in this solution is Bright Cluster Manager. The installation of this software is made by using a bare-metal method and plugging a bright bootable virtual disk to the head node. The software image used for the installation of the other nodes is a modified version of the auto-generated software image of bright (Red-Hat 7.9).

```
# cmsh
# softwareimage
# show default-image
# clone default-image test-image
# # modify test-image properties with set <properties> [value]
#commit
```



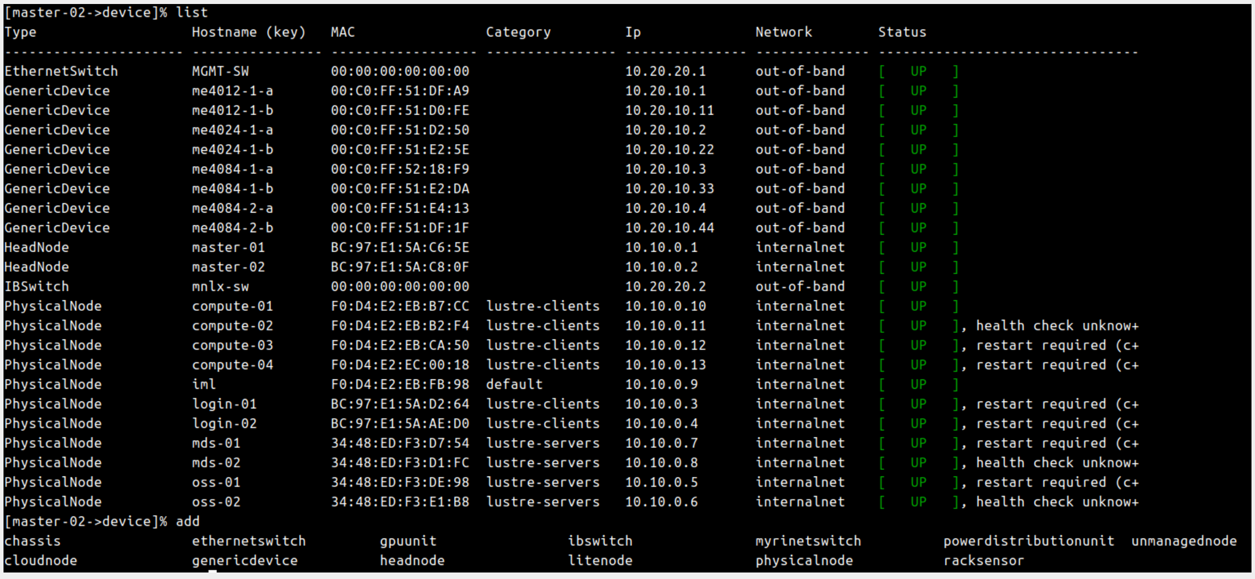
```
[master-02->softwareimage]% list
Name (key)          Path                                     Kernel version                     Nodes
------------------- ---------------------------------------- ---------------------------------- --------
default-image       /cm/images/default-image                 3.10.0-1160.el7.x86_64             0
login_image         /cm/images/login_image                   3.10.0-1160.el7.x86_64             2
lustre_client_image /cm/images/lustre_client_image           3.10.0-1160.el7.x86_64             5
lustre_image        /cm/images/lustre_image                  3.10.0-1160.2.1.el7_lustre.x86_64  4
```



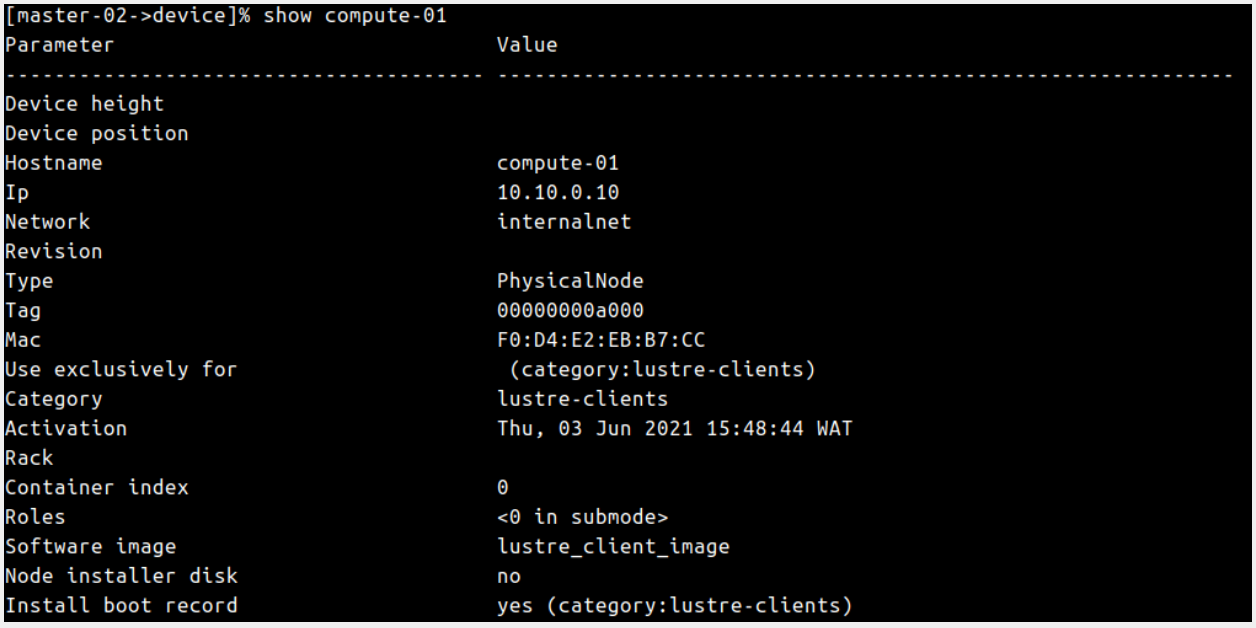
```
[master-02->softwareimage]% show lustre_image
Parameter                        Value
-------------------------------- ------------------------------------------------
Name                             lustre_image
Nodes                            4
Revision
Path                             /cm/images/lustre_image
Creation time                    Fri, 11 Jun 2021 12:51:12 WAT
Kernel version                   3.10.0-1160.2.1.el7_lustre.x86_64
Kernel parameters
Kernel output console            tty0
Kernel modules                   <57 in submode>
Enable SOL                       no
SOL Port                         ttyS1
SOL Speed                        115200
SOL Flow Control                 yes
FSPart                           /cm/images/lustre_image
Boot FSPart                      /cm/images/lustre_image/boot
Notes                            <0B>
```

Add nodes to cluster

```
# cmsh
# device
# add physicalnode test
##nodetypes:chassis,ethernetswitch,gpuunit,ibswitch,
myrinetswitch,powerdistributionunit,unmanagednode,cloudnode,
genericdevice, headnode, litenode, physicalnode, racksensor
# show test#
```

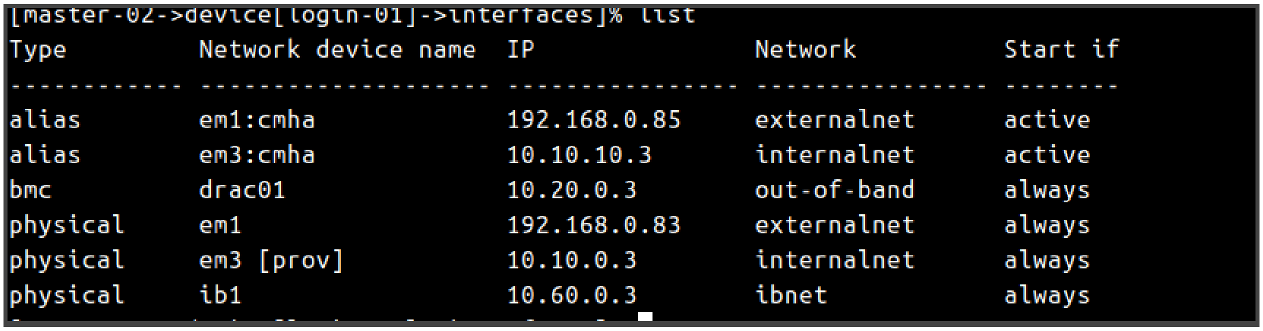


```
[master-02->device]% list
Type                   Hostname (key)   MAC                Category         Ip              Network        Status
---------------------- ---------------- ------------------ ---------------- --------------- -------------- --------------------------------
EthernetSwitch         MGMT-SW          00:00:00:00:00:00                   10.20.20.1      out-of-band    [   UP   ]
GenericDevice          me4012-1-a       00:C0:FF:51:DF:A9                   10.20.10.1      out-of-band    [   UP   ]
GenericDevice          me4012-1-b       00:C0:FF:51:D0:FE                   10.20.10.11     out-of-band    [   UP   ]
GenericDevice          me4024-1-a       00:C0:FF:51:D2:50                   10.20.10.2      out-of-band    [   UP   ]
GenericDevice          me4024-1-b       00:C0:FF:51:E2:5E                   10.20.10.22     out-of-band    [   UP   ]
GenericDevice          me4084-1-a       00:C0:FF:52:18:F9                   10.20.10.3      out-of-band    [   UP   ]
GenericDevice          me4084-1-b       00:C0:FF:51:E2:DA                   10.20.10.33     out-of-band    [   UP   ]
GenericDevice          me4084-2-a       00:C0:FF:51:E4:13                   10.20.10.4      out-of-band    [   UP   ]
GenericDevice          me4084-2-b       00:C0:FF:51:DF:1F                   10.20.10.44     out-of-band    [   UP   ]
HeadNode               master-01        BC:97:E1:5A:C6:5E                   10.10.0.1       internalnet    [   UP   ]
HeadNode               master-02        BC:97:E1:5A:C8:0F                   10.10.0.2       internalnet    [   UP   ]
IBSwitch               mnlx-sw          00:00:00:00:00:00                   10.20.20.2      out-of-band    [   UP   ]
PhysicalNode           compute-01       F0:D4:E2:EB:B7:CC  lustre-clients   10.10.0.10      internalnet    [   UP   ]
PhysicalNode           compute-02       F0:D4:E2:EB:B2:F4  lustre-clients   10.10.0.11      internalnet    [   UP   ], health check unknow+
PhysicalNode           compute-03       F0:D4:E2:EB:CA:50  lustre-clients   10.10.0.12      internalnet    [   UP   ], restart required (c+
PhysicalNode           compute-04       F0:D4:E2:EC:00:18  lustre-clients   10.10.0.13      internalnet    [   UP   ], restart required (c+
PhysicalNode           iml              F0:D4:E2:EB:FB:98  default          10.10.0.9       internalnet    [   UP   ]
PhysicalNode           login-01         BC:97:E1:5A:D2:64  lustre-clients   10.10.0.3       internalnet    [   UP   ], restart required (c+
PhysicalNode           login-02         BC:97:E1:5A:AE:D0  lustre-clients   10.10.0.4       internalnet    [   UP   ], restart required (c+
PhysicalNode           mds-01           34:48:ED:F3:D7:54  lustre-servers   10.10.0.7       internalnet    [   UP   ], restart required (c+
PhysicalNode           mds-02           34:48:ED:F3:D1:FC  lustre-servers   10.10.0.8       internalnet    [   UP   ], health check unknow+
PhysicalNode           oss-01           34:48:ED:F3:DE:98  lustre-servers   10.10.0.5       internalnet    [   UP   ], restart required (c+
PhysicalNode           oss-02           34:48:ED:F3:E1:B8  lustre-servers   10.10.0.6       internalnet    [   UP   ], health check unknow+
[master-02->device]% add
chassis                ethernetswitch         gpuunit                ibswitch               myrinetswitch          powerdistributionunit  unmanagednode
cloudnode              genericdevice          headnode               litenode               physicalnode           racksensor
```

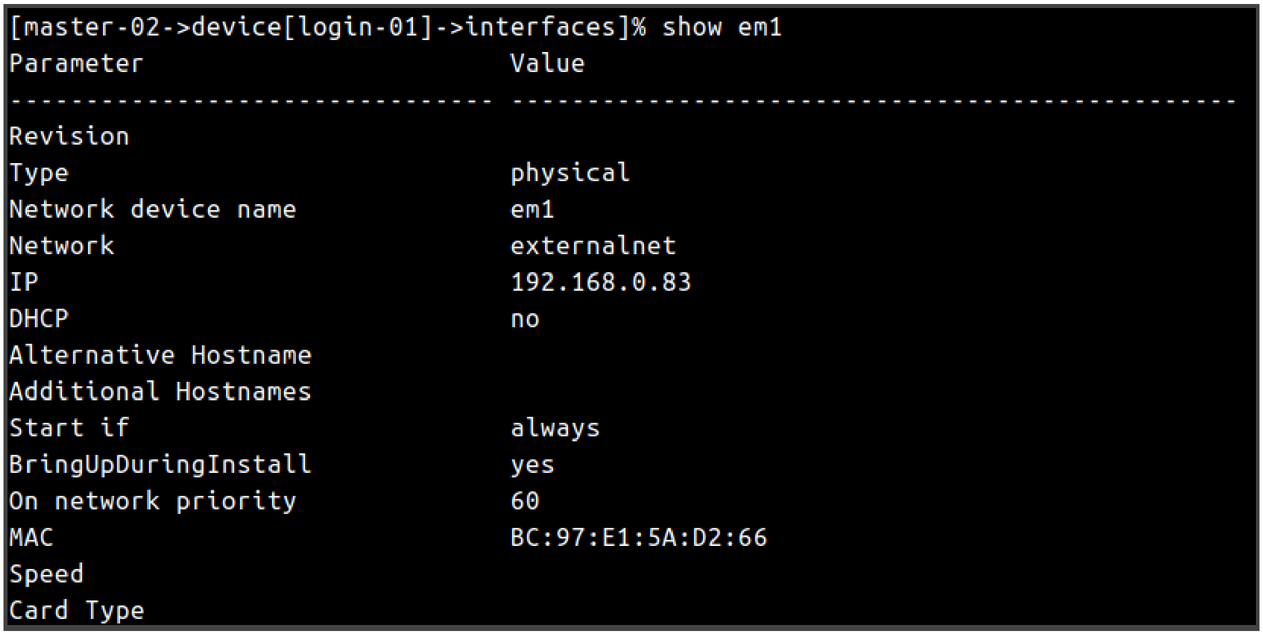


```
[master-02->device]% show compute-01
Parameter                               Value
--------------------------------------- ------------------------------------------------------------
Device height
Device position
Hostname                                compute-01
Ip                                      10.10.0.10
Network                                 internalnet
Revision
Type                                    PhysicalNode
Tag                                     00000000a000
Mac                                     F0:D4:E2:EB:B7:CC
Use exclusively for                      (category:lustre-clients)
Category                                lustre-clients
Activation                              Thu, 03 Jun 2021 15:48:44 WAT
Rack
Container index                         0
Roles                                   <0 in submode>
Software image                          lustre_client_image
Node installer disk                     no
Install boot record                     yes (category:lustre-clients)
```

```
# modify test properties with set <properties> [value]
# some values that should be changed: management network; interfaces; provisioning interface; installmode=auto; installbootrecord=yes
# interfaces test
# add physical [name]
# # interfaces types : alias      bmc        bond       bridge     netmap     physical   tunnel     vlan
# show [name]
# # modify [name] interface properties with set <properties> [value]
#commit
```



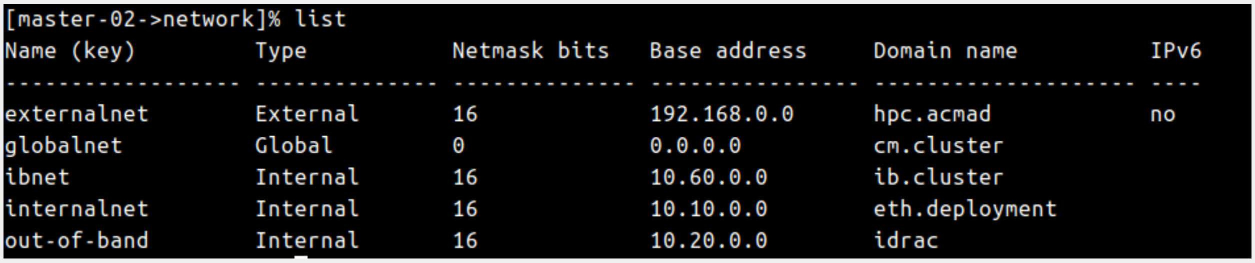
```
[master-02->device[login-01]->interfaces]% list
Type         Network device name  IP               Network          Start if
------------ -------------------- ---------------- ---------------- --------
alias        em1:cmha             192.168.0.85     externalnet      active
alias        em3:cmha             10.10.10.3       internalnet      active
bmc          drac01               10.20.0.3        out-of-band      always
physical     em1                  192.168.0.83     externalnet      always
physical     em3 [prov]           10.10.0.3        internalnet      always
physical     ib1                  10.60.0.3        ibnet            always
```



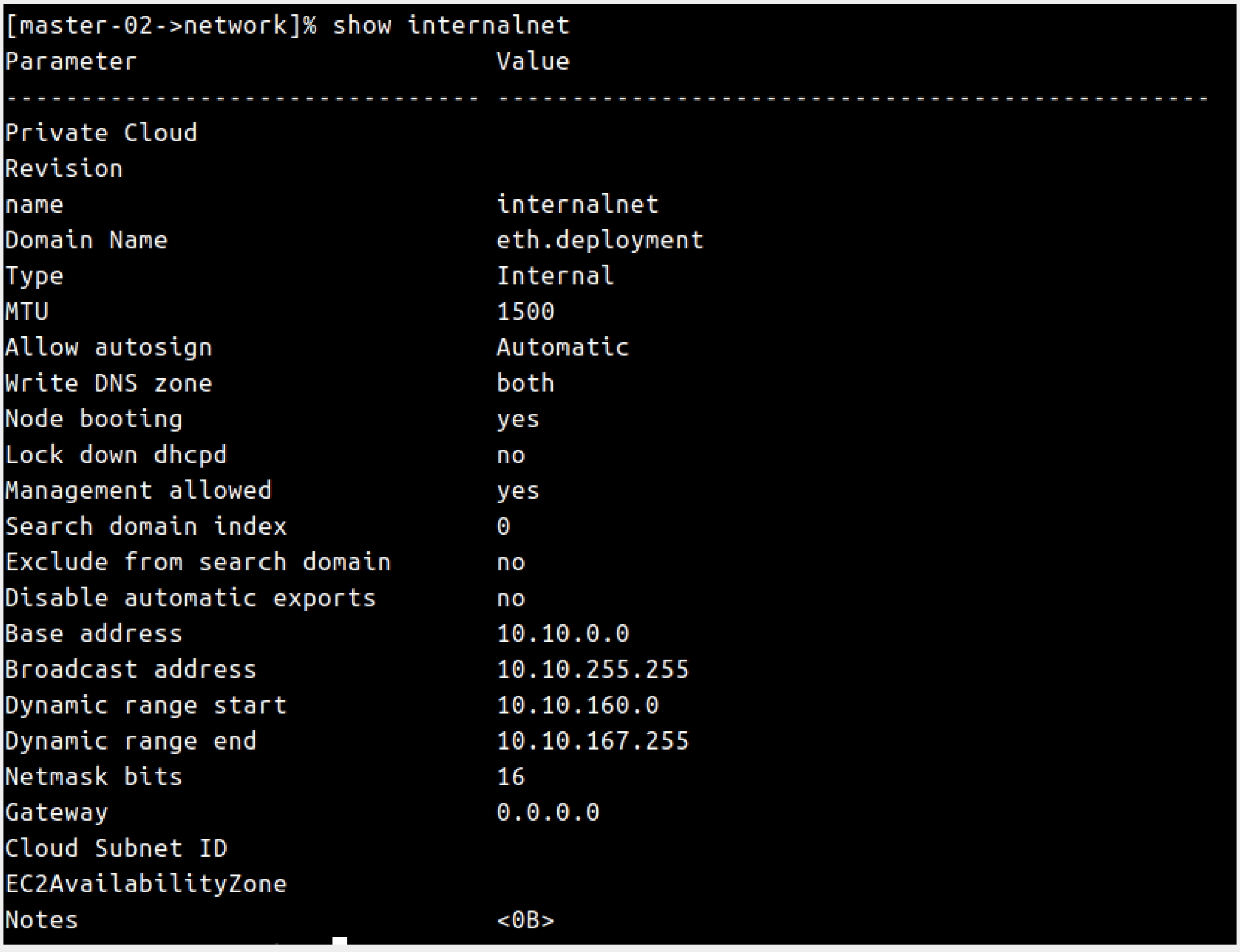
```
[master-02->device[login-01]->interfaces]% show em1
Parameter                        Value
-------------------------------- ------------------------------------------------
Revision
Type                             physical
Network device name              em1
Network                          externalnet
IP                               192.168.0.83
DHCP                             no
Alternative Hostname
Additional Hostnames
Start if                         always
BringUpDuringInstall             yes
On network priority              60
MAC                              BC:97:E1:5A:D2:66
Speed
Card Type
```

Add network to cluster

```
# cmsh
# network
# add [name]
# modify network properties with set <properties> [value]
#commit
```



```
[master-02->network]% list
Name (key)        Type          Netmask bits  Base address    Domain name         IPv6
----------------- ------------- ------------- --------------- ------------------- ----
externalnet       External      16            192.168.0.0     hpc.acmad           no
globalnet         Global        0             0.0.0.0         cm.cluster
ibnet             Internal      16            10.60.0.0       ib.cluster
internalnet       Internal      16            10.10.0.0       eth.deployment
out-of-band       Internal      16            10.20.0.0       idrac
```



```
[master-02->network]% show internalnet
Parameter                        Value
-------------------------------- ------------------------------------------------
Private Cloud
Revision
name                             internalnet
Domain Name                      eth.deployment
Type                             Internal
MTU                              1500
Allow autosign                   Automatic
Write DNS zone                   both
Node booting                     yes
Lock down dhcpd                  no
Management allowed               yes
Search domain index              0
Exclude from search domain       no
Disable automatic exports        no
Base address                     10.10.0.0
Broadcast address                10.10.255.255
Dynamic range start              10.10.160.0
Dynamic range end                10.10.167.255
Netmask bits                     16
Gateway                          0.0.0.0
Cloud Subnet ID
EC2AvailabilityZone
Notes                            <0B>
```

In band management

```
#pdsh -g all 'date' | dshbak -c
#commit
```

Out of band management

```
#ipmitool -H 10.20.0.2 -U [user] -P '[passwd]' -I lanplus power status
```

Login servers

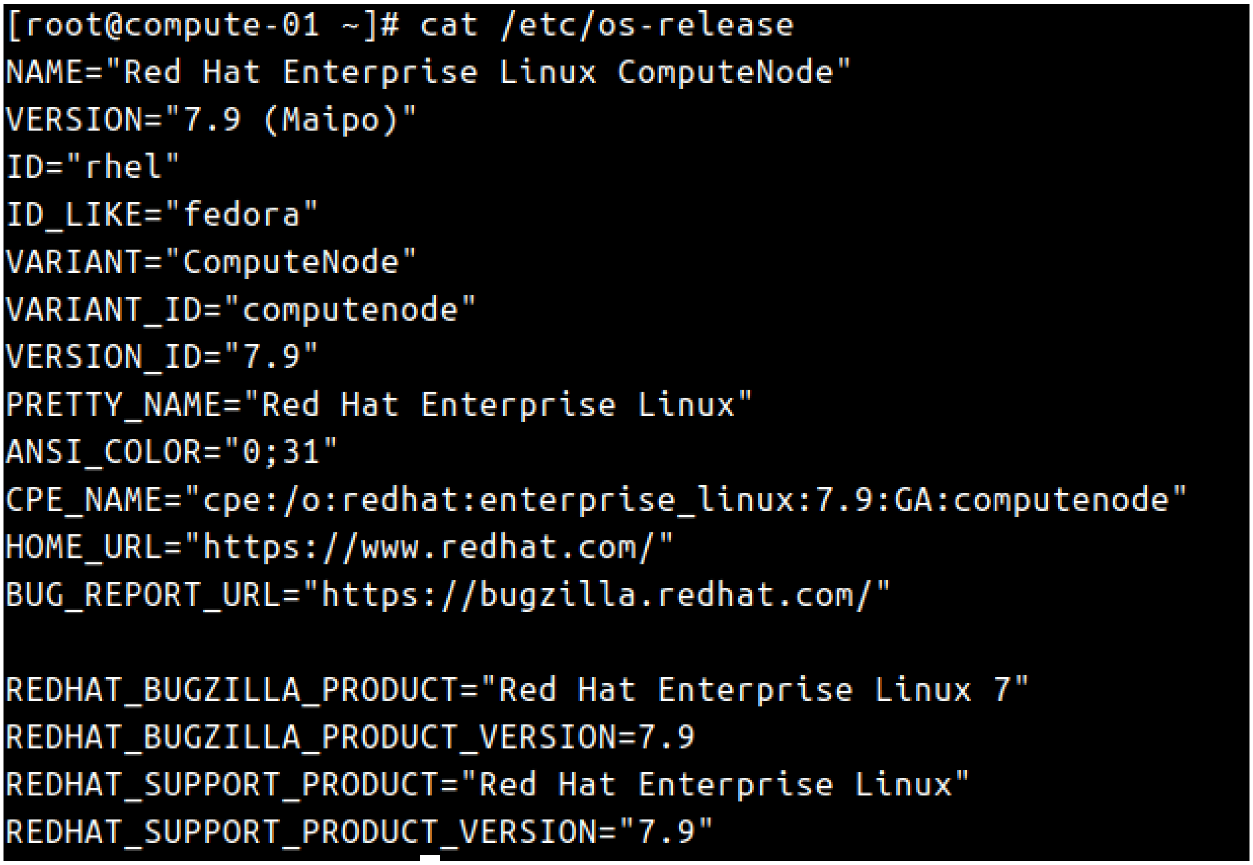
Login servers allow SSH access to the scientific users. Once the users have been logged in, they have at their disposal:

- 4 computing nodes, whose resources are allocatable by means of Slurm queue system
- Disk space on home
- Disk space on Lustre filesystem

Access to the active login server is granted on IP 192.168.0.83, by means of SSH only.

Computing servers

The computing servers allow the users to run computational workflows. Computing nodes are not connected on external VLAN, and their operating system has IP on VLAN 100 only. Therefore, access on computing nodes is granted only after a login on the login servers.
IP addresses of the four computing servers span from 10.10.0.10 to 10.10.0.13.
On computing nodes is applied a diverse Red Hat operating system (Red Hat Enterprise Linux ComputeNode) from the one offered by cluster manager Bright (Red Hat Enterprise Linux Server). This is made to match the system purpose of the compute node with the license Red Hat bought for that node. The following pictures shows the details



```
[root@compute-01 ~]# cat /etc/os-release
NAME="Red Hat Enterprise Linux ComputeNode"
VERSION="7.9 (Maipo)"
ID="rhel"
ID_LIKE="fedora"
VARIANT="ComputeNode"
VARIANT_ID="computenode"
VERSION_ID="7.9"
PRETTY_NAME="Red Hat Enterprise Linux"
ANSI_COLOR="0;31"
CPE_NAME="cpe:/o:redhat:enterprise_linux:7.9:GA:computenode"
HOME_URL="https://www.redhat.com/"
BUG_REPORT_URL="https://bugzilla.redhat.com/"

REDHAT_BUGZILLA_PRODUCT="Red Hat Enterprise Linux 7"
REDHAT_BUGZILLA_PRODUCT_VERSION=7.9
REDHAT_SUPPORT_PRODUCT="Red Hat Enterprise Linux"
REDHAT_SUPPORT_PRODUCT_VERSION="7.9"
```

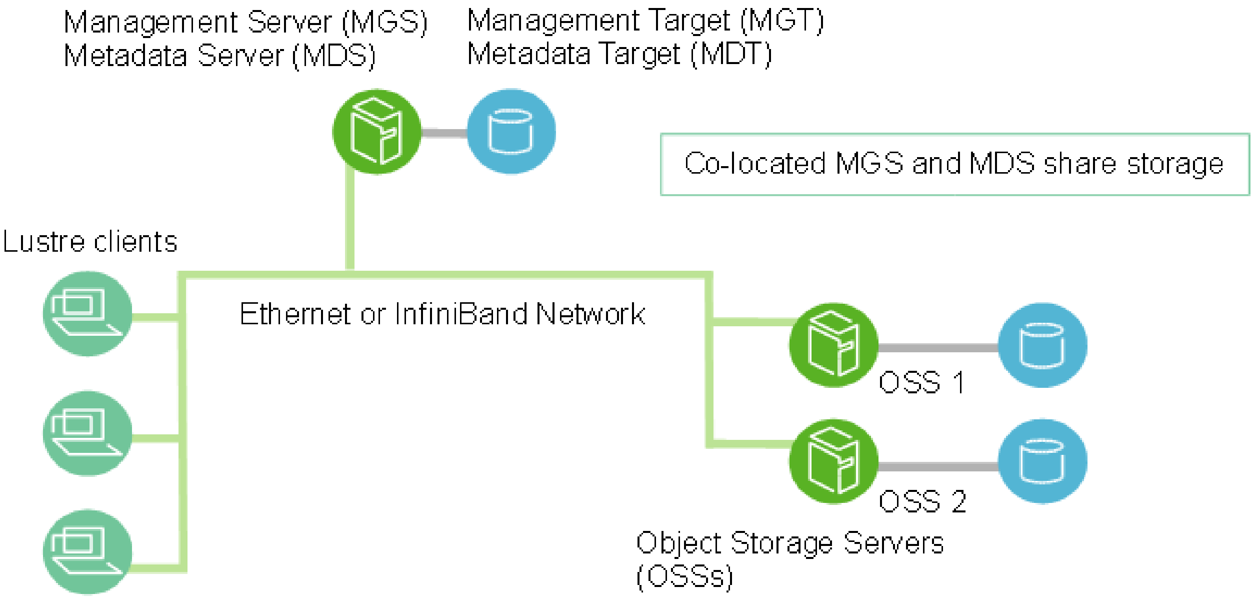
Lustre servers

The four Lustre servers (mds-01, mds-02, oss-01, oss-02) have been installed with RedHat 7.9 and configured with Lustre 2.12.6. This release is freely downloadable from https://downloads.whamcloud.com/public/lustre/latest-2.12-release/.
Note that, at the moment of this writing, the compatibility matrix for Lustre request a RedHat 7.9 server: https://wiki.whamcloud.com/display/PUB/Lustre+Support+Matrix.

The main components of a Lustre filesystem are:

- The mds servers, that stores the metadata in a MDT, a metadata target, and makes it available to the Lustre clients
- The oss servers, that store the data in one or more OST, and handle the network requests from the Lustre clients
- The Lustre clients, mounting the Lustre filesystem

To acquire further information about the Lustre filesystem please visit the official documentation.



In the ACMAD HPC Cluster the used targets for mds server and oss server are provided by the disk arrays:

- The mds servers are connected to the ME4024 disk array, that provides the metadata target (a single RAID10 composed by 10 SSD disks). The mds servers are configured in active/standby, so only the active server can mount the target.
- The oss servers are connected to the two ME4084 disk arrays, that provides a total of 16 OSTs (object storage targets). 8 OSTs are mounted on oss-01, while 8 OSTs are mounted on oss-02. Each OST is a RAID6 of 10 disks.

The Lustre servers require a special kernel to be installed in place of the stock one. For Lustre 2.12.6, the following RPM kernel has been installed on the Lustre servers:

kerrnel-3.10.0-1160.2.1.el7_lustre.x86_64.rpm

The Lustre clients in the ACMAD HPC Cluster are the master servers, the login servers, the computing nodes and the monitoring server. These servers don’t need a special kernel to mount the Lustre client, but a few RPMs are enough to have the filesystem mounted (e.g. the Lustre patchless client).

The Lustre client service is provided by means of a patchless
No ofed, stock drivers
Metadata servers
Object storage servers
Multipath to equally distribute OSTs
Client mount point
Failover capabilities
TCP fallback
/etc/modprobe.d/lustre.conf

Lustre filesystem deployment

The following are command lines used to deploy the Lustre filesystem.

Create mgs

The mgs service has been created on the RAID10 target published by the ME4024 disk array.
The mgt target doesn’t need performance, since it’s a configuration target only.
The mgt target will be made available on both the mds-01 and mds-02 servers (only one at a time), and on both the InfiniBand and Ethernet network.

```
mkfs.lustre --mgs --servicenode 10.10.0.7@tcp,10.60.0.7@o2ib --servicenode 10.10.0.8@tcp,10.60.0.8@o2ib --backfstype=ldiskfs /dev/mapper/mgt
```

Create mdt

The mgs service has been created on the RAID10 target published by the ME4024 disk array.
The mdt target needs maximum performance. This target will be managed by the active mds server, with failover capability on the standby mds server.

```
mkfs.lustre   --mdt  --fsname lustre --mgsnode 10.10.0.7@tcp,10.60.0.7@o2ib --mgsnode 10.10.0.8@tcp,10.60.0.8@o2ib --servicenode 10.10.0.7@tcp,10.60.0.7@o2ib --servicenode 10.10.0.8@tcp,10.60.0.8@o2ib --index 0 /dev/mapper/mdt
```

Create ost

The OSTs have been created on the two ME4084 disk arrays. These 16 volumes are visible on both the oss nodes, but can be mounted on a single oss. Both the oss servers are active, and each mount a total of 8 OSTs: 4 OSTs from the first ME4084, 4 OSTs from the other ME4084.

```
for i in {01..16}
do
mkfs.lustre --ost \
--fsname lustre \
--index $i \
--mgsnode 10.10.0.7@tcp,10.60.0.7@o2ib \
--mgsnode 10.10.0.8@tcp,10.60.0.8@o2ib \
--servicenode 10.10.0.5@tcp,10.60.0.5@o2ib \
--servicenode 10.10.0.6@tcp,10.60.0.6@o2ib \
--backfstype=ldiskfs \
/dev/mapper/ost$i
mkdir -p /lustre/ost/$i
done
```

How to start the Lustre filesystem

The Lustre filesystem requests an order for start of services.

1. start mgt on a mds (mds-01 is the primary)

   ```
   mount -t lustre /dev/mapper/mgt /lustre/mgt
   ```

2. start mdt on a mds (mds-01 is the primary)

   ```
   mount -t lustre /dev/mapper/mdt /lustre/mdt0
   ```

3. start ost on oss-01

   ```
   for i in {01..16..2}
   do
           mount -t lustre  /dev/mapper/ost$i /lustre/ost/$i
   done
   ```

4. Start ost on oss-02

   ```
   for i in {02..16..2}
   do
           mount -t lustre  /dev/mapper/ost$i /lustre/ost/$i
   done
   ```

5. Start the clients (to run from the master server)

   ```
   pdsh -g lustre-clients mount -t lustre 10.60.0.7@o2ib,10.60.0.8@o2ib0:/lustre /lustre/
   ```

How to stop the Lustre filesystem

An order to stop the Lustre filesystem is requested.

1. stop lustre client everywhere

   ```
   pdsh -g lustre-clients umount /lustre
   ```

2. stop ost (umount of all the osts)

   ```
   pdsh -w oss-01 umount -t lustre -a
   pdsh -w oss-02 umount -t lustre -a
   ```

3. stop mdt (umount mdt)

   ```
   pdsh -w mds-01 umount /lustre/mdt0
   ```

4. stop mgt (umount mgt)

   ```
   pdsh -w mds-01 umount /lustre/mgt
   ```

5. Lustre_rmmod

   ```
   pdsh -g all lustre_rmmod
   ```

Lustre network - LNet

The Lustre network kernel module has been configured on all the servers to use InfiniBand as primary network layer for transport of I/O. The Ethernet channel has been configured as a fallback, in case InfiniBand is unavailable.

Configuration of lustre network: /etc/modprobe.d/lustre.conf

```
compute-[01-04],mds-[01-02],oss-[01-02],iml
----------------
options lnet networks="o2ib0(ib0),tcp(em1)"
----------------
master-[01-02]
----------------
options lnet networks="o2ib0(ib1),tcp(eno3)"

----------------
login-[01-02]
----------------
options lnet networks="o2ib0(ib1),tcp(em3)"
```

Storage controller configuration on ME4 series

The storage controller on all the ME4 series disk arrays have been configured mainly via command line. We report here the basic operations used to configure the disk arrays.
All the disk arrays have a redundant controller, and each controller has a management interface that the system administrator can use to login on the controller.

Here’s the procedure for configuring the ME4 systems with pools, hosts, and volumes.

Associating hosts and initiators

Each disk array is connected to two hosts. On each disk array the hosts have been configured to reflect their original hostname (therefore the ME4012 sees two hosts, master-01 and master-02.

a. On the operating system of the servers: check their initiators with lsscsi
   `lsscsi --host --transport`
   Then in the ME4 cli you will see the initiators for recognize them
   `Show initiators`
b. Next the initiators will need a nickname to make every node recognizable
   `set initiator id 00aaa0b000cccc00 nickname host-number-one-1`

c. Then the hosts can be created using their respective initiators
```
create host initiators 00aaa0b000cccc00,11ddd1e000fff000
host-number-one
```

Creating pools

To create the pools with the disks you'll need to choose the raid type.

a. First we can see the disks so you can choose which of those to use
```
Show disks
```
b. Then you create a pool with the vdisk command
```
create vdisk level raid6 disks 0.0-6 spare 0.7,0.8 vdisk01
```
When configuring a raid10 or raid 50 the command is different, and you need to specify all the mirrors
```
create vdisk level raid10 disks 0.0-1:0.2-3:0.4-5:0.6-7:0.8-9
spare 0.10,0.11 vdisk01
```
c. For adding spares globally use the command add spares
```
add spares 0.80-83
```
d. Then you can see all the pools and their details
```
Show pools
```

Creating volumes

Finally, after creating the pools, we can create the volumes.

a. you can create a volume with the create volume command specifying the pool
```
create volume vdisk vdisk01 size 2TB vol01
```
b. Then you can see our created volumes
```
show volumes
```
c. There is also the possibility to expand the volumes, in this way you can even set all the free space available to a single volume if needed
```
expand volume size 4TB vol01
expand volume size max vol01
```

Mapping the volumes and hosts

After creating the volumes you need to map them to the hosts that will use them.

a. First we must check the volumes and ports

   ```
   show volumes
   show ports
   ```

b. Now we can create the map using the map command, you need to specify the ports and the lun that will be used

   ```
   map volume vol01 access rw initiator 00aaa0b000cccc00,11ddd1e111fff111,22ggg2h222iii22,33jjj3k333lll333 ports a0,a1,a2,a3 lun 0
   ```

c. You can unmap a volume if needed, for example in the case you need to change some details and create another
   unmap volume initiator

   ```
   00aaa0b000cccc00,11ddd1e111fff111,22ggg2h222iii22,33jjj3k333lll333 vol01,vol02
   ```

If you need all the command list with their explanations here's the link for the DELL guide: *Dell ME4 commands guide*.

Queue system

The jobs that are going to be executed in future should be managed via Workload Manager. In this cluster the Workload manager that is used is Slurm. To configure Slurm, Bright cluster manager offers his help with different tools and commands managed by his daemon (CMDaemon). To configure from scratch a workload manager bright offers NCursers (blue console) management:

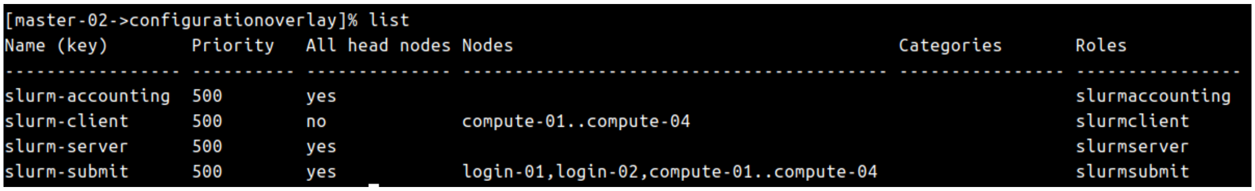
```
#cm-wlm-setup
```

As the configuration is made from scratch we have to be sure that there is no other workload manager configured by default so in the pop-up is chosen *disable.* Just after that, is created a new instance by selecting *Setup (Step-by-step)*
During the setup are chosen the roles of any node so in the following table is described the configuration of slurm:

Type	*Name (key)*	*Server nodes*	*Submit nodes*	*Client nodes*
Slurm	*slurm*	*master-01,master-02*	*login-01,login-02,master-01,master-02,compute-[01-04]*	*compute-01..compute-04*

To update the configuration of slurm (priority etc) and check the configuration, the following commands are used:

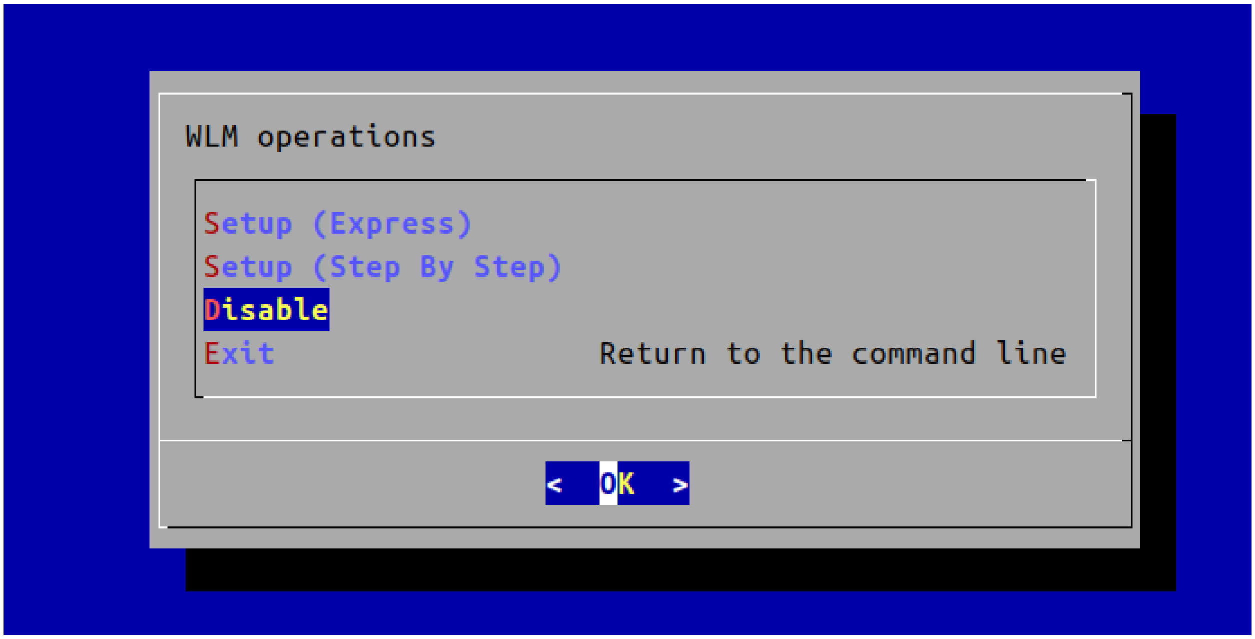
```
#cmsh >
#wlm
#list
#configurationoverlay
#list
```



```
[master-02->configurationoverlay]% list
Name (key)        Priority   All head nodes Nodes                                    Categories       Roles
----------------- ---------- -------------- ---------------------------------------- ---------------- ----------------
slurm-accounting  500        yes                                                                      slurmaccounting
slurm-client      500        no             compute-01..compute-04                                    slurmclient
slurm-server      500        yes                                                                      slurmserver
slurm-submit      500        yes            login-01,login-02,compute-01..compute-04                  slurmsubmit
```

To disable slurm the following command is used and disable is chosen:

```
Cm-wlm-setup
```



Power off/on procedures

In order to execute a graceful power off or power on of the HPC Cluster, please adopt the following procedures.

Power off

1. Ensure no users are running on the infrastructure

Run the `squeue` command to check the Slurm queue system is empty. If jobs are running, you may want to decide to terminate them (`scancel JOBID`).
Be sure all the jobs are drained from the queue system.

2. Stop the queue system

Check the queue system has been stopped server and client side.

3. Umount network filesystem everywhere

Network filesystem must be unmounted everywhere before a proper shutdown.

On the login and compute nodes, please umount /home, /cm/shared and /lustre (be sure no users are using the filesystems).

```
umount /home
umount /cm/shared
umount /lustre
lustre_rmmod
```

On the master nodes, if mounted, umount /lustre.

```
umount /lustre
lustre_rmmod
```

4. Shutdown Lustre on Lustre servers

The Lustre servers read and write from the ME disk arrays. For a proper shutdown, the Lustre targets need to be unmounted. Please run the following commands in the right order on all the Lustre servers.

```
# on oss-01 and oss-02
umount /home
umount /cm/shared
umount -t lustre -a

# on mds-01 and mds-02
umount /home
umount /cm/shared
umount /lustre/mdt0
umount /lustre/mgt
```

5. Stop monitoring services

Login on the IML and stop the services.

```
service omd stop
```

6. Shutdown all the servers

Once services have been stopped and the network filesystems unmounted, you can proceed in server shutdown.

```
# computing and login nodes
pdsh -g lustre-clients poweroff

# lustre servers
pdsh -g lustre-servers poweroff

# monitoring IML server
ssh iml poweroff
```

7. Shutdown the passive masternode

From the active master nodes, shutdown the passive master node. Be sure network filesystems are unmounted.

8. Final actions

From the active masternode, check all the servers have been shut down.

```
[root@master-02 ~]# cmsh -c “device list”
```

From the active masternode, stop the NFS service and unmount the /home and /cm/shared.

```
[root@master-02 ~]# cmsh -c “device use master; services; stop nfs”
[root@master-02 ~]# cmsh -c “device use master-01; services; stop nfs”
[root@master-02 ~]# umount /home
[root@master-02 ~]# umount /cm/shared
```

In order to guarantee a proper shutdown of the storage controllers on the disks array, you can connect to their management interface and shutdown their controllers.

```
[root@master-02 ~]# ssh me4012-1-a “shutdown both”
[root@master-02 ~]# ssh me4024-1-a “shutdown both”
[root@master-02 ~]# ssh me4084-1-a “shutdown both”
[root@master-02 ~]# ssh me4084-2-a “shutdown both”
```

9. Shutdown the active masternode

Shutdown the active masternode.

Power on

In order to execute power on of the HPC Cluster, please adopt the below procedure with the following important note:

Master-02 is the only server to turn on manually. All the other servers are going to turn on via out-of-band network (idrac) that is accessible from master-02.

1. Restart master-02 manually

To be completed on-site.

2. Restart passive masternode

```
[root@master-02 ~]# cmsh -c “device use master-01; power on”
```

3. Restart all the storages

Login to each storage controller and check they are UP.

```
[root@master-02 ~]# ssh administrator@me4012-1-a “show
shutdown-status”
Storage Controller A up

Storage Controller B up

Other MC Status Operational
```

If not, the command to restart them is the following (must be executed on each ME disk array).

```
[root@master-02 ~]# ssh administrator@me4012-1-a “restart both”
```

Check the /home and /cm/shared are mounted on the active masternode, and start the NFS service.

```
[root@master-02 ~]# cmsh -c “device use master; services; start nfs”
[root@master-02 ~]# cmsh -c “device use master-01; services; start
nfs”
```

4. Restart lustre servers

```
[root@master-02 ~]# cmsh -c “device; foreach -n
oss-[01-02],mds-[01-02] (power on)”
```

Wait a few minutes until they are all reachable. Once reachable, load the lustre modules

```
[root@master-02 ~]# pdsh -g lustre-servers modprobe lustre
```

To start the lustre filesystem please follow “ How to start the Lustre filesystem” in this manual.

5. Restart lustre clients (Login-[01-02], Compute-nodes[01-04]) and mount /lustre on master nodes

```
[root@master-02 ~]# cmsh -c “device; foreach -n
login-[01-02],compute-[01-04] (power on)”
```

Wait a few minutes until they are all reachable. Once reachable, load the lustre modules

```
[root@master-02 ~]# pdsh -g lustre-clients modprobe lustre
```

And mount the lustre filesystem

```
[root@master-02 ~]# pdsh -g lustre-clients mount /lustre
```

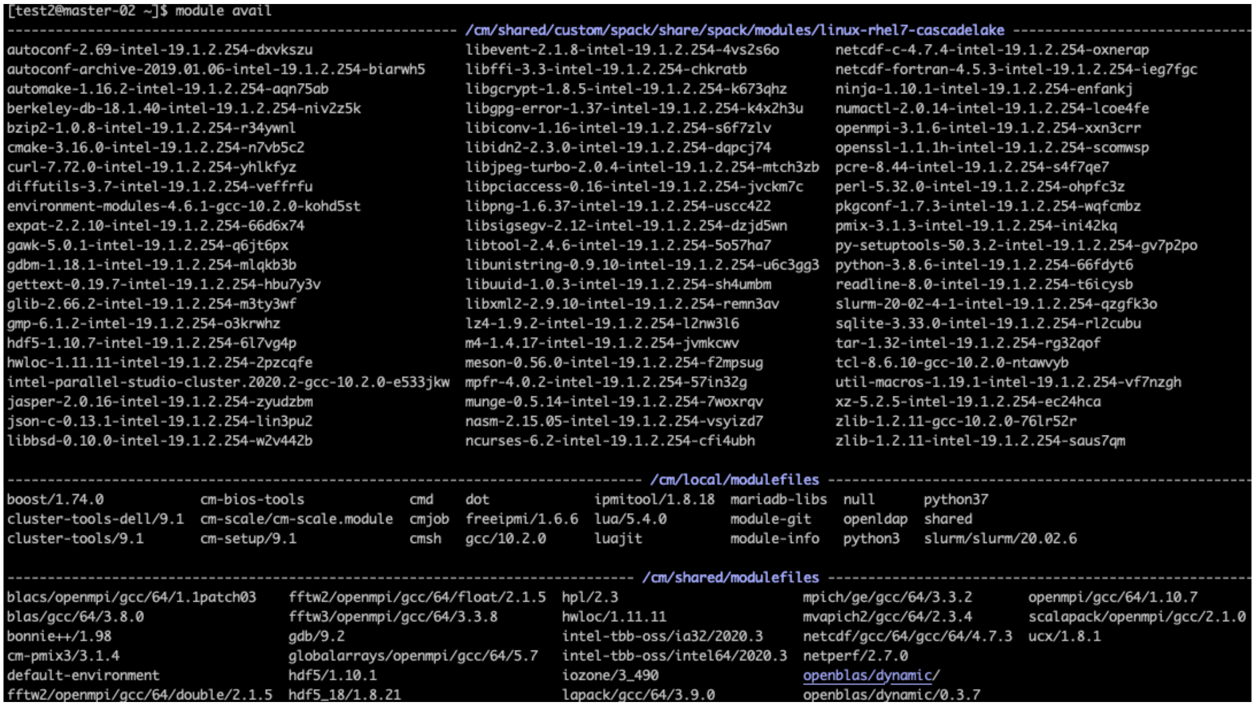
```
[root@master-02 ~]# pdsh -g headnode modprobe lustre
[root@master-02 ~]# pdsh -g headnode mount -t lustre
10.60.0.7@o2ib,10.60.0.8@o2ib0:/lustre /lustre/
```

6. Restart monitoring server

```
[root@master-02 ~]# cmsh -c ‘device use iml; power on’
```

Scientific software

Scientific software has been installed in the /cm/shared filesystem on the master nodes, managed by Bright CM and highly available to all the login and computing nodes.
Scientific libraries, compilers and application are made available to the users by means of environment-modules (http://modules.sourceforge.net/), pre-installed by Bright CM.
To have a view on the installed software, run “module avail”.



```
[test2@master-02 ~]$ module avail
--------------------------------------------------------- /cm/shared/custom/spack/share/spack/modules/linux-rhel7-cascadelake ------------------------------
autoconf-2.69-intel-19.1.2.254-dxvkszu                  libevent-2.1.8-intel-19.1.2.254-4vs2s6o      netcdf-c-4.7.4-intel-19.1.2.254-oxnerap
autoconf-archive-2019.01.06-intel-19.1.2.254-biarwh5    libffi-3.3-intel-19.1.2.254-chkratb          netcdf-fortran-4.5.3-intel-19.1.2.254-ieg7fgc
automake-1.16.2-intel-19.1.2.254-aqn75ab                libgcrypt-1.8.5-intel-19.1.2.254-k673qhz     ninja-1.10.1-intel-19.1.2.254-enfankj
berkeley-db-18.1.40-intel-19.1.2.254-niv2z5k            libgpg-error-1.37-intel-19.1.2.254-k4x2h3u   numactl-2.0.14-intel-19.1.2.254-lcoe4fe
bzip2-1.0.8-intel-19.1.2.254-r34ywnl                    libiconv-1.16-intel-19.1.2.254-s6f7zlv       openmpi-3.1.6-intel-19.1.2.254-xxn3crr
cmake-3.16.0-intel-19.1.2.254-n7vb5c2                   libidn2-2.3.0-intel-19.1.2.254-dqpcj74       openssl-1.1.1h-intel-19.1.2.254-scomwsp
curl-7.72.0-intel-19.1.2.254-yhlkfyz                    libjpeg-turbo-2.0.4-intel-19.1.2.254-mtch3zb pcre-8.44-intel-19.1.2.254-s4f7qe7
diffutils-3.7-intel-19.1.2.254-veffrfu                  libpciaccess-0.16-intel-19.1.2.254-jvckm7c   perl-5.32.0-intel-19.1.2.254-ohpfc3z
environment-modules-4.6.1-gcc-10.2.0-kohd5st            libpng-1.6.37-intel-19.1.2.254-uscc422       pkgconf-1.7.3-intel-19.1.2.254-wqfcmbz
expat-2.2.10-intel-19.1.2.254-66d6x74                   libsigsegv-2.12-intel-19.1.2.254-dzjd5wn     pmix-3.1.3-intel-19.1.2.254-ini42kq
gawk-5.0.1-intel-19.1.2.254-q6jt6px                     libtool-2.4.6-intel-19.1.2.254-5o57ha7       py-setuptools-50.3.2-intel-19.1.2.254-gv7p2po
gdbm-1.18.1-intel-19.1.2.254-mlqkb3b                    libunistring-0.9.10-intel-19.1.2.254-u6c3gg3 python-3.8.6-intel-19.1.2.254-66fdyt6
gettext-0.19.7-intel-19.1.2.254-hbu7y3v                 libuuid-1.0.3-intel-19.1.2.254-sh4umbm       readline-8.0-intel-19.1.2.254-t6icysb
glib-2.66.2-intel-19.1.2.254-m3ty3wf                    libxml2-2.9.10-intel-19.1.2.254-remn3av      slurm-20-02-4-1-intel-19.1.2.254-qzgfk3o
gmp-6.1.2-intel-19.1.2.254-o3krwhz                      lz4-1.9.2-intel-19.1.2.254-l2nw3l6           sqlite-3.33.0-intel-19.1.2.254-rl2cubu
hdf5-1.10.7-intel-19.1.2.254-6l7vg4p                    m4-1.4.17-intel-19.1.2.254-jvmkcwv           tar-1.32-intel-19.1.2.254-rg32qof
hwloc-1.11.11-intel-19.1.2.254-2pzcqfe                  meson-0.56.0-intel-19.1.2.254-f2mpsug        tcl-8.6.10-gcc-10.2.0-ntawvyb
intel-parallel-studio-cluster.2020.2-gcc-10.2.0-e533jkw mpfr-4.0.2-intel-19.1.2.254-57in32g          util-macros-1.19.1-intel-19.1.2.254-vf7nzgh
jasper-2.0.16-intel-19.1.2.254-zyudzbm                  munge-0.5.14-intel-19.1.2.254-7woxrqv        xz-5.2.5-intel-19.1.2.254-ec24hca
json-c-0.13.1-intel-19.1.2.254-lin3pu2                  nasm-2.15.05-intel-19.1.2.254-vsyizd7        zlib-1.2.11-gcc-10.2.0-76lr52r
libbsd-0.10.0-intel-19.1.2.254-w2v442b                  ncurses-6.2-intel-19.1.2.254-cfi4ubh         zlib-1.2.11-intel-19.1.2.254-saus7qm

--------------------------------------------------------------------------------- /cm/local/modulefiles ------------------------------------------------------
boost/1.74.0           cm-bios-tools            cmd   dot            ipmitool/1.8.18 mariadb-libs null     python37
cluster-tools-dell/9.1 cm-scale/cm-scale.module cmjob freeipmi/1.6.6 lua/5.4.0       module-git   openldap shared
cluster-tools/9.1      cm-setup/9.1             cmsh  gcc/10.2.0     luajit          module-info  python3  slurm/slurm/20.02.6

-------------------------------------------------------------------------------- /cm/shared/modulefiles -------------------------------------------------------
blacs/openmpi/gcc/64/1.1patch03  fftw2/openmpi/gcc/64/float/2.1.5 hpl/2.3                      mpich/ge/gcc/64/3.3.2      openmpi/gcc/64/1.10.7
blas/gcc/64/3.8.0                fftw3/openmpi/gcc/64/3.3.8       hwloc/1.11.11                mvapich2/gcc/64/2.3.4      scalapack/openmpi/gcc/2.1.0
bonnie++/1.98                    gdb/9.2                          intel-tbb-oss/ia32/2020.3    netcdf/gcc/64/gcc/64/4.7.3 ucx/1.8.1
cm-pmix3/3.1.4                   globalarrays/openmpi/gcc/64/5.7  intel-tbb-oss/intel64/2020.3 netperf/2.7.0
default-environment              hdf5/1.10.1                      iozone/3_490                 openblas/dynamic/
fftw2/openmpi/gcc/64/double/2.1.5 hdf5_18/1.8.21                  lapack/gcc/64/3.9.0          openblas/dynamic/0.3.7
```

Software and modules are provided by Spack (https://spack.io/), a tool that automatically downloads and builds the dependency chain needed by a software.

Intel Compiler

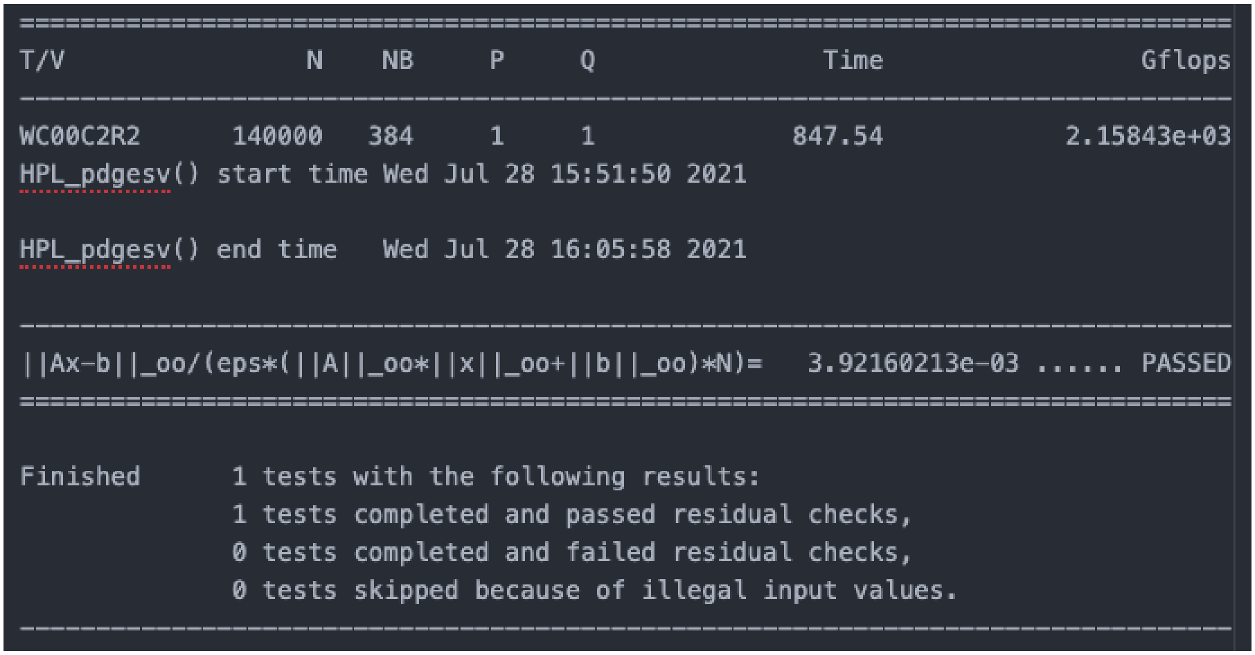
The Intel Compiler has been installed by Spack, and registered with the purchased license. The license is readable in /cm/shared/custom/spack/etc/spack/licenses/intel/license.lic.

HPL

The HPL Linpack benchmark (http://www.netlib.org/linpack/) is available in the home of the test2 user.
To build it, just load the necessary modules and enter the right folder. Then, run the build.sh wrapper.

```
[test2@login-01 ~]$ module load openmpi-3.1.6-intel-19.1.2.254-xxn3crr
[test2@login-01 ~]$ module load
intel-parallel-studio-cluster.2020.2-gcc-10.2.0-e533jkw
[test2@login-01 ~]$ cd HPL/mp_linpack/
[test2@login-01 mp_linpack]$ ./build.sh
```

In order to maximize the memory usage (the compute nodes have 192GB of RAM) the Linpack has been run on a compute node with Ns 140000 and Block Size 384, that fill the memory at 80%. With these values we got 2.15 Tflops on the two CPUs of the compute node. This means about 8.6 Tflops using 4 computing nodes (the theoretic peak is 12Tflops, so this means a 71% efficiency).



```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WC00C2R2      140000   384     1     1             847.54             2.15843e+03
HPL_pdgesv() start time Wed Jul 28 15:51:50 2021

HPL_pdgesv() end time   Wed Jul 28 16:05:58 2021

--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=   3.92160213e-03 ...... PASSED
================================================================================

Finished      1 tests with the following results:
              1 tests completed and passed residual checks,
              0 tests completed and failed residual checks,
              0 tests skipped because of illegal input values.
--------------------------------------------------------------------------------
```

WRF

WRF and WPS have been installed with version 4.3 (WRF: https://github.com/wrf-model/WRF/archive/refs/tags/v4.3.tar.gz, WPS: https://github.com/wrf-model/WPS/archive/refs/tags/v4.3.tar.gz).

To build them, the following software stack has been built and used:

```
module load \
intel-parallel-studio-cluster.2020.2-gcc-10.2.0-e533jkw \
netcdf-fortran-4.5.3-intel-19.1.2.254-ieg7fgc \
openmpi-3.1.6-intel-19.1.2.254-xxn3crr \
hdf5-1.10.7-intel-19.1.2.254-6l7vg4p \
jasper-2.0.16-intel-19.1.2.254-zyudzbm \
libpng-1.6.37-intel-19.1.2.254-uscc422  \
zlib-1.2.11-intel-19.1.2.254-saus7qm
```

The WRF and WPS have been built in the `/cm/shared/custom/software/ folder`.

Monitoring

The tool that is used in this cluster for event monitoring is Check MK (https://checkmk.com).
This section explains with a step-by-step method the installation procedure of server and client nodes as well as the configuration of the software on this cluster
To access the monitoring website it is necessary to create a tunnel to internal network following the below command:

```
ssh 192.168.0.85 -AL 1575:10.10.0.9:443 -L 5900:10.10.0.9:5900
```

Just after that launch at your browser to reach the monitoring website:
http://127.0.0.1:1575
Please ask the administrator for the credentials to login.

Server Install

1) Download the .rpm package of check mk server from official website on server node. In our case the download is made on IML node
 https://checkmk.com/download?edition=cfe&version=stable

```
#wget -c
https://download.checkmk.com/checkmk/2.0.0p5/check-mk-raw-2.0.0p5-el7-38.x86_64.rpm
```

2) install the package

```
#yum install $PATH/check-mk-raw-2.0.0p5-el7-38.x86_64.rpm
```

3) Check the installation

```
#which omd
```

4) Create a site

```
#omd create [name] ##(for example Acmad)
```

!!!! Take note of user credentials that are generated: The credentials can be changed via check_mk website!!!

5) Start the site

```
#omd start [name] ##(for example Acmad)
```

6) Check the site: check creates a filesystem *tmpfs* on */run/user/*

```
#df -h
```

7) Access the webpage via browser:

http://IP/[nomesite] ##(for example Acmad)

Client Install

1) Download the .rpm package of check client from check_mk server installed on IML node

```
#wget -c http://10.10.0.9/acmad/agents
```

2) Install the package

```
#pdsh -g check-mk-clients yum install
/cm/shared/check-mk-agent.noarch.rpm
```

3) Check the installation

```
#rpm -aq | grep check-mk-agent.noarch.rpm
#which check_mk_agent
```

4) Install *xinetd*

```
#pdsh -g check-mk-clients yum -y --disablerepo=*
--enablerepo=local-base install xinetd
```

5) Check *xinetd* functionality & check_mk configuration on it

```
#systemctl status xinetd
#ls /etc/xinetd | grep -i check_mk
```

6) Check the site: check creates a filesystem tmpfs on */run/user/*

```
#df -h
```

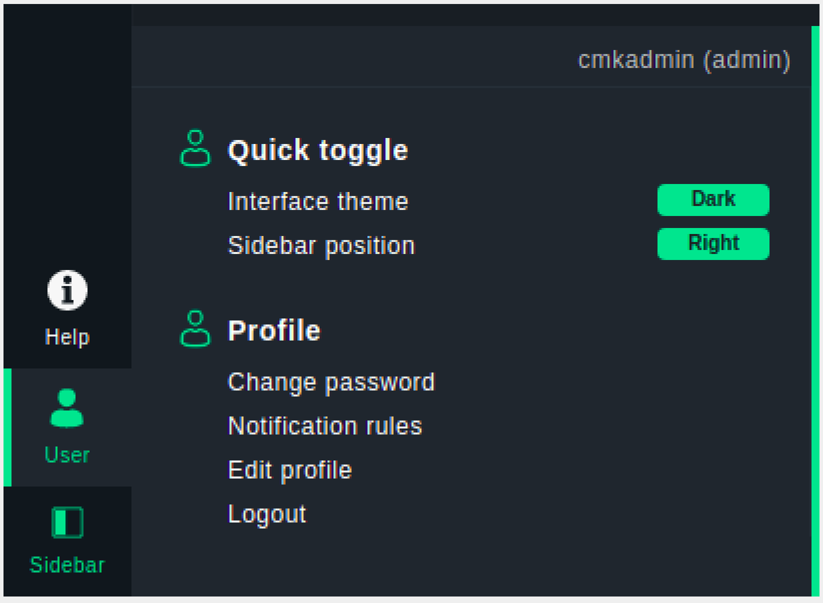
Configuration of Check MK: Adding nodes & services

To access the check MK monitoring website the following link can be used from the internal network (10.10.0.0/16):
http://10.10.0.9/acmad
Please ask the administrator for the credentials to login.

All the following operations are made on check_mk webpage that is placed on IML node.

1) Change the password offered by *omd create* during server installation

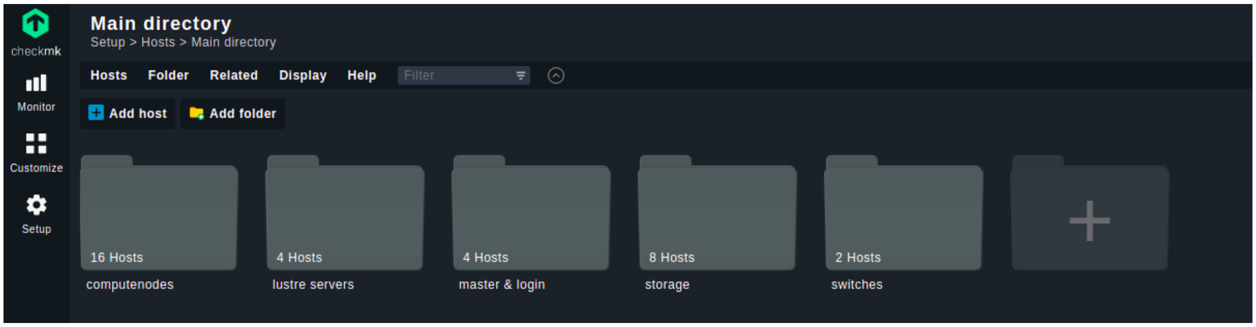
```
LEFT PANEL: User>Change Password
```

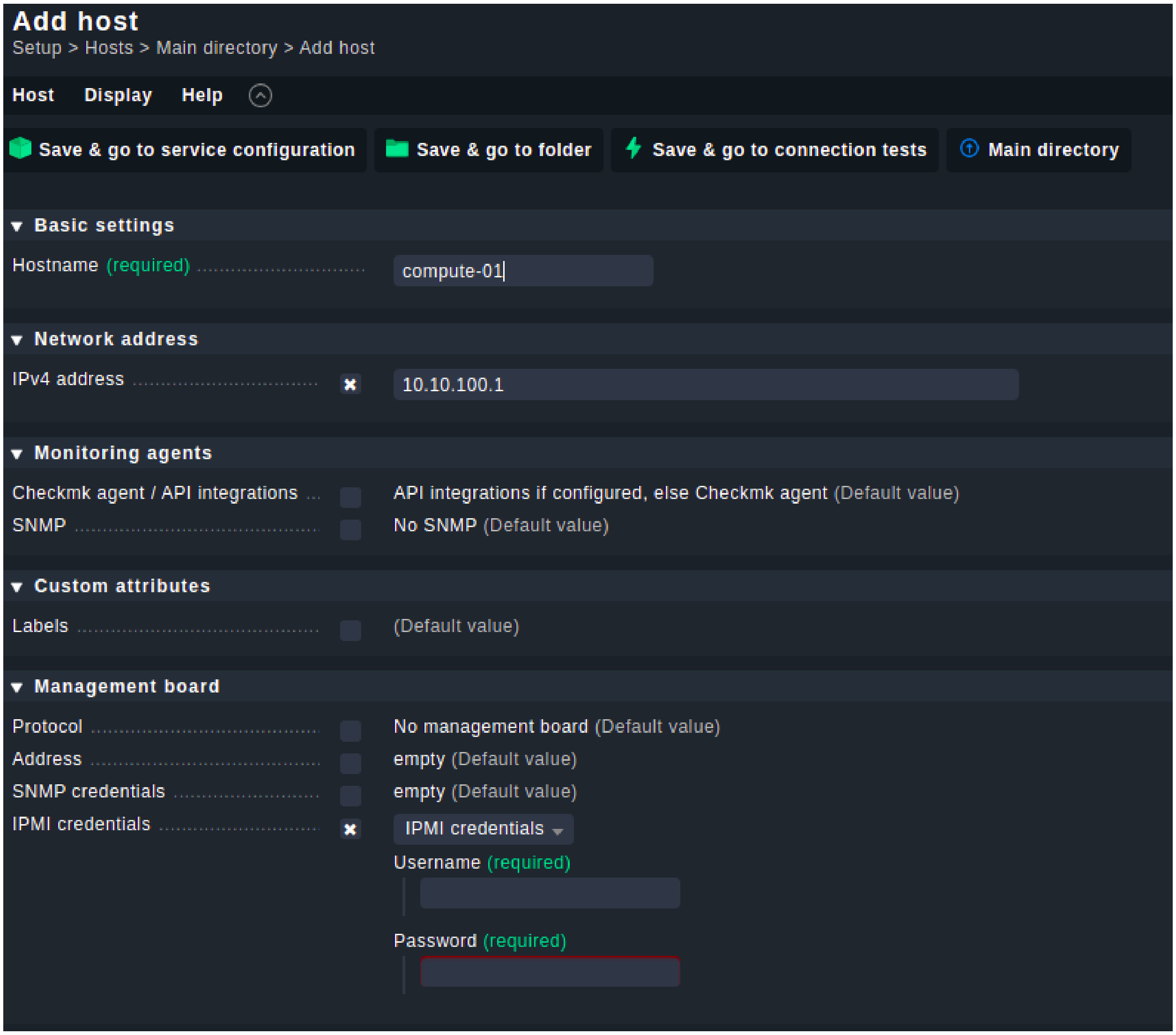


Interface theme: Dark
Sidebar position: Right

2) Add the client/agent nodes to check_mk

```
LEFT PANEL: SetUp>Hosts>add node ##specify Node Name; IP;  IPMI
credentials
```




Add host
Setup > Hosts > Main directory > Add host
Host
Display
Help
Save & go to service configuration
Save & go to folder
Save & go to connection tests
Main directory

Basic settings

Hostname (required) compute-01

Network address

IPv4 address .... ☒ 10.10.100.1

Monitoring agents

Checkmk agent / API integrations ... □ API integrations if configured, else Checkmk agent (Default value)

SNMP .... □ No SNMP (Default value)

Custom attributes

Labels .... □ (Default value)

Management board

Protocol .... □ No management board (Default value)

Address .... □ empty (Default value)

SNMP credentials .... □ empty (Default value)

IPMI credentials .... ☒ IPMI credentials

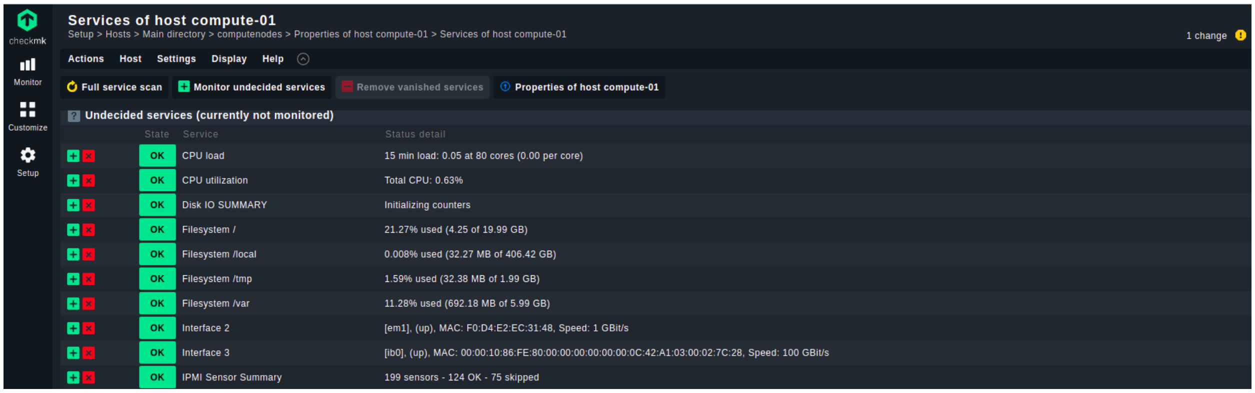
Username (required)

Password (required)

3) Edit monitored services & save changes

## Add the monitored services

```
LEFT PANEL: SetUp>Hosts>[select a node]> Save & go to service
configuration > Select the service that are going to be monitored &
the services that are not (Hardware parameters is recommended to be
under investigation such as CPU, RAM ecc)
```



Services of host compute-01

Setup > Hosts > Main directory > computenodes > Properties of host compute-01 > Services of host compute-01

Undecided services (currently not monitored)

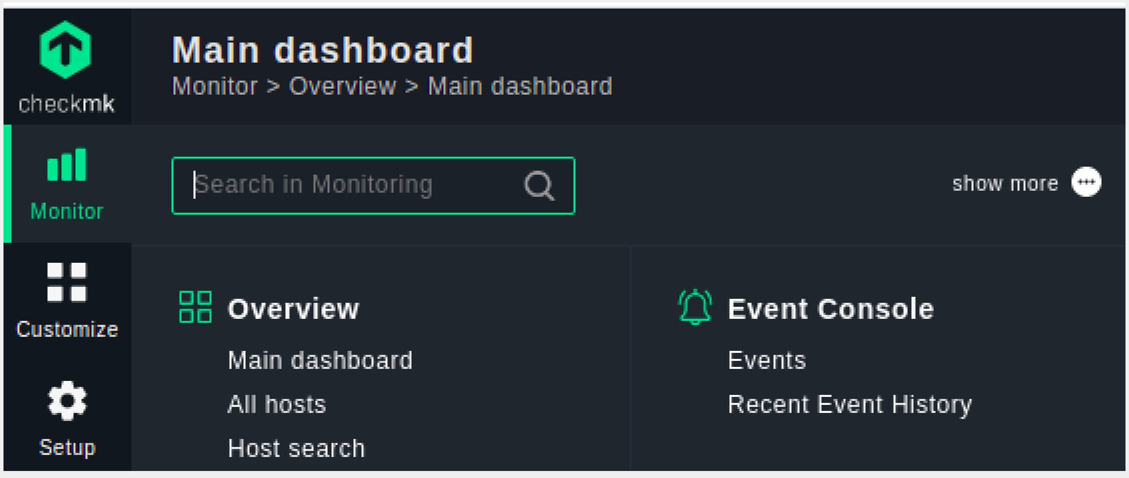
State	Service	Status detail
OK	CPU load	15 min load: 0.05 at 80 cores (0.00 per core)
OK	CPU utilization	Total CPU: 0.63%
OK	Disk IO SUMMARY	Initializing counters
OK	Filesystem /	21.27% used (4.25 of 19.99 GB)
OK	Filesystem /local	0.008% used (32.27 MB of 406.42 GB)
OK	Filesystem /tmp	1.59% used (32.38 MB of 1.99 GB)
OK	Filesystem /var	11.28% used (692.18 MB of 5.99 GB)
OK	Interface 2	[em1], (up), MAC: F0:D4:E2:EC:31:48, Speed: 1 GBit/s
OK	Interface 3	[ib0], (up), MAC: 00:00:10:86:FE:80:00:00:00:00:00:00:0C:42:A1:03:00:02:7C:28, Speed: 100 GBit/s
OK	IPMI Sensor Summary	199 sensors - 124 OK - 75 skipped

## Save changes

```
At the top right corner select changes than click on “Activate on
selected sites”
```

4) Check nodes health

```
LEFT PANEL:Monitor>All hosts
## each service monitored in a node could be in 5 states:
    -  OK
    -  WARNING
    -  UNCRITICAL
    -  CRITICAL
    -  PENDING
```



checkmk

Main dashboard

Monitor > Overview > Main dashboard

Search in Monitoring

show more

Monitor

Customize

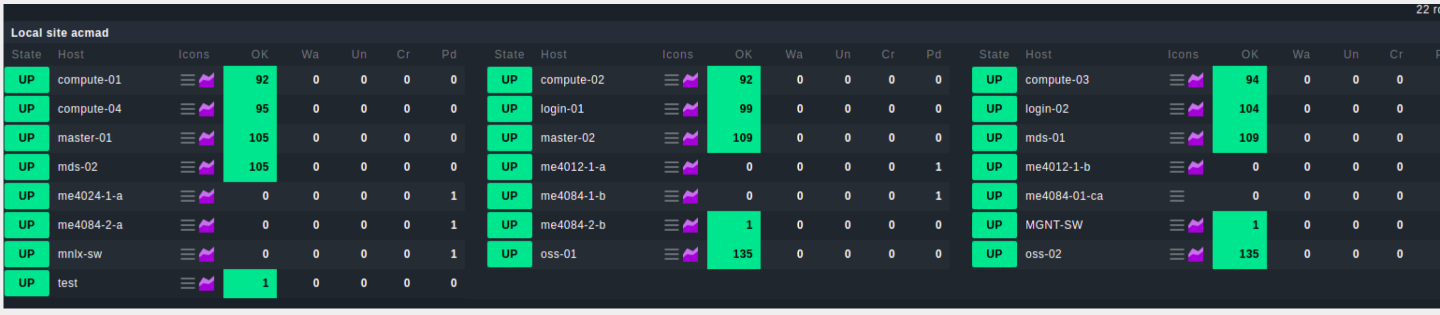
Setup

Overview

- Main dashboard
- All hosts
- Host search

Event Console

- Events
- Recent Event History



22 ro

Local site acmad

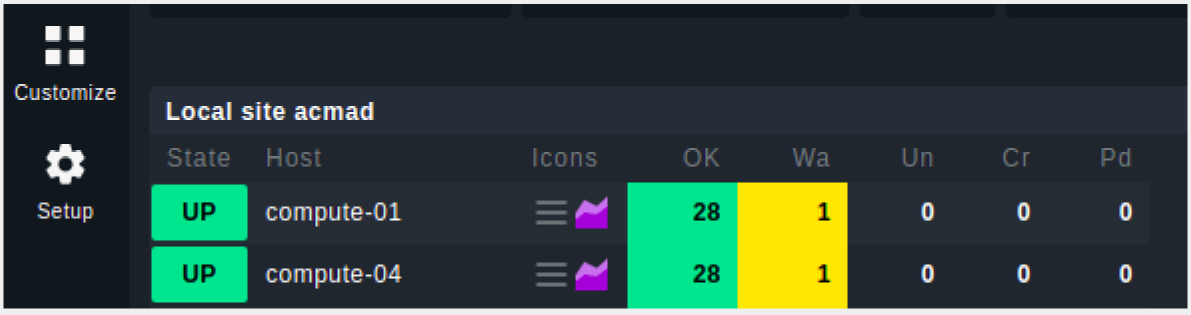
State	Host	Icons	OK	Wa	Un	Cr	Pd
UP	compute-01	☰ 📈	92	0	0	0	0
UP	compute-04	☰ 📈	95	0	0	0	0
UP	master-01	☰ 📈	105	0	0	0	0
UP	mds-02	☰ 📈	105	0	0	0	0
UP	me4024-1-a	☰ 📈	0	0	0	0	1
UP	me4084-2-a	☰ 📈	0	0	0	0	1
UP	mnlx-sw	☰ 📈	0	0	0	0	1
UP	test	☰ 📈	1	0	0	0	0

State	Host	Icons	OK	Wa	Un	Cr	Pd
UP	compute-02	☰ 📈	92	0	0	0	0
UP	login-01	☰ 📈	99	0	0	0	0
UP	master-02	☰ 📈	109	0	0	0	0
UP	me4012-1-a	☰ 📈	0	0	0	0	1
UP	me4084-1-b	☰ 📈	0	0	0	0	1
UP	me4084-2-b	☰ 📈	1	0	0	0	0
UP	oss-01	☰ 📈	135	0	0	0	0

State	Host	Icons	OK	Wa	Un	Cr
UP	compute-03	☰ 📈	94	0	0	0
UP	login-02	☰ 📈	104	0	0	0
UP	mds-01	☰ 📈	109	0	0	0
UP	me4012-1-b	☰ 📈	0	0	0	0
UP	me4084-01-ca	☰	0	0	0	0
UP	MGNT-SW	☰ 📈	1	0	0	0
UP	oss-02	☰ 📈	135	0	0	0

4) Check services health

ALL HOSTS DASHBOARD: Select a state in a node (Ok, Warning, Uncritical, Critical, Pending) to see which monitored service is in that certain state



Customize

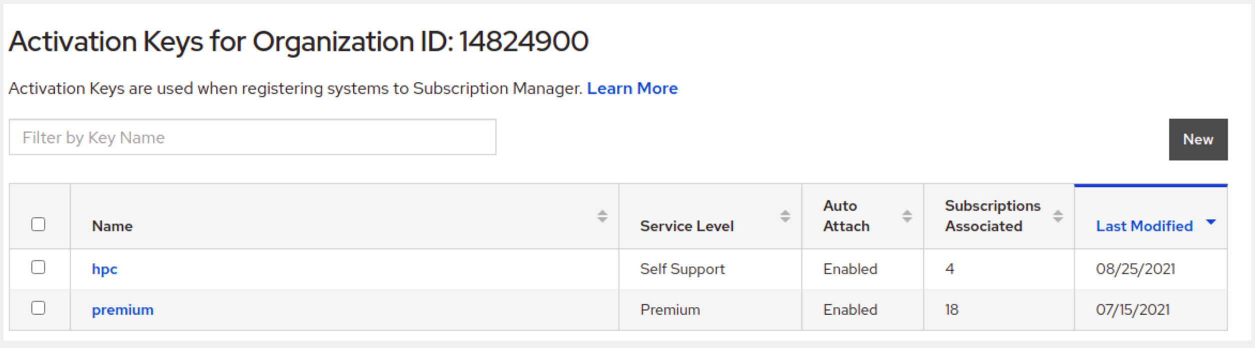
Setup

Local site acmad

State	Host	Icons	OK	Wa	Un	Cr	Pd
UP	compute-01	☰ 📈	28	1	0	0	0
UP	compute-04	☰ 📈	28	1	0	0	0

Red Hat subscription

To activate the subscription on systems that are installed with Red Hat 7.9 (Red Hat Enterprise Linux Server,Red Hat Enterprise Linux ComputeNode) is used the activation key method.
On Red Hat portal is created two categories of subscriptions, *premium* for server nodes installed with Red Hat Enterprise Linux Server and *hpc* for compute nodes installed with Red Hat Enterprise Linux ComputeNode.

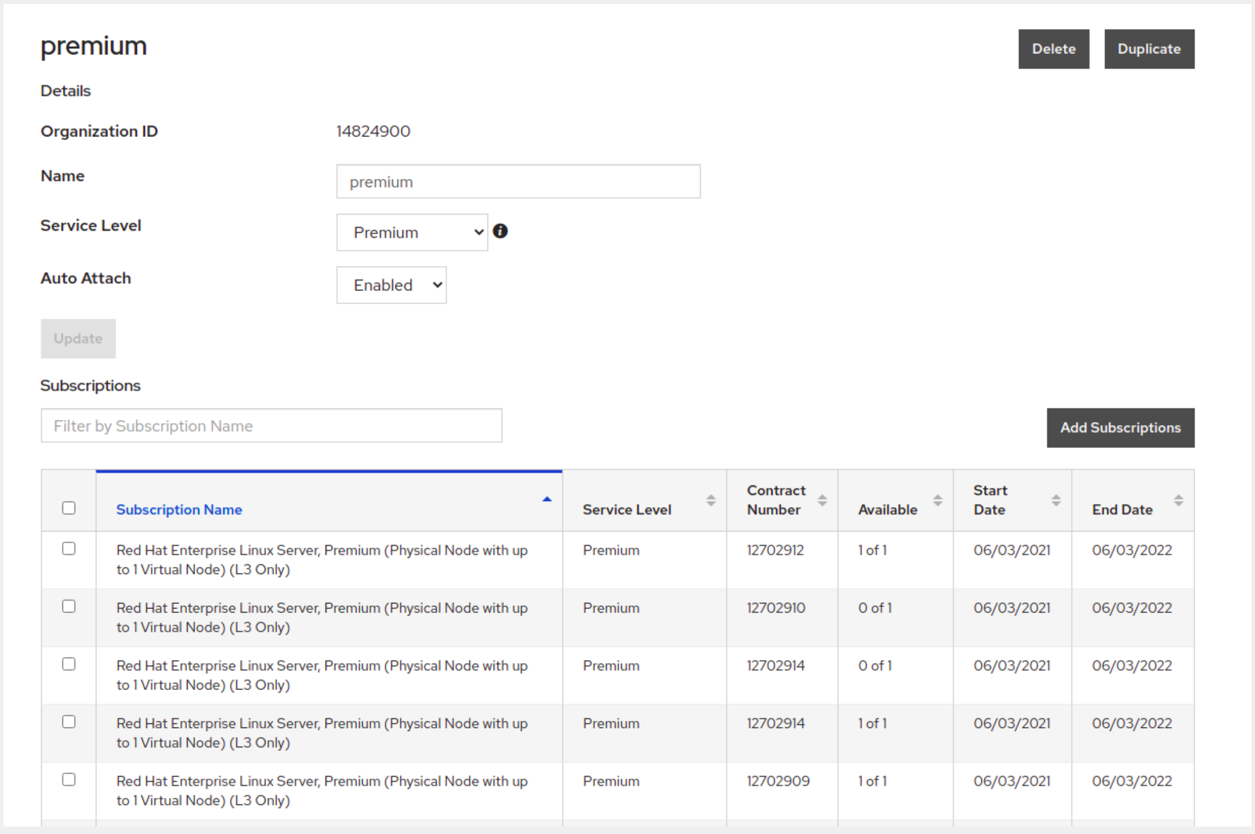


Activation Keys for Organization ID: 14824900

Activation Keys are used when registering systems to Subscription Manager. Learn More

Filter by Key Name

□	Name	Service Level	Auto Attach	Subscriptions Associated	Last Modified
□	hpc	Self Support	Enabled	4	08/25/2021
□	premium	Premium	Enabled	18	07/15/2021



premium

Details

Organization ID 14824900

Name: premium

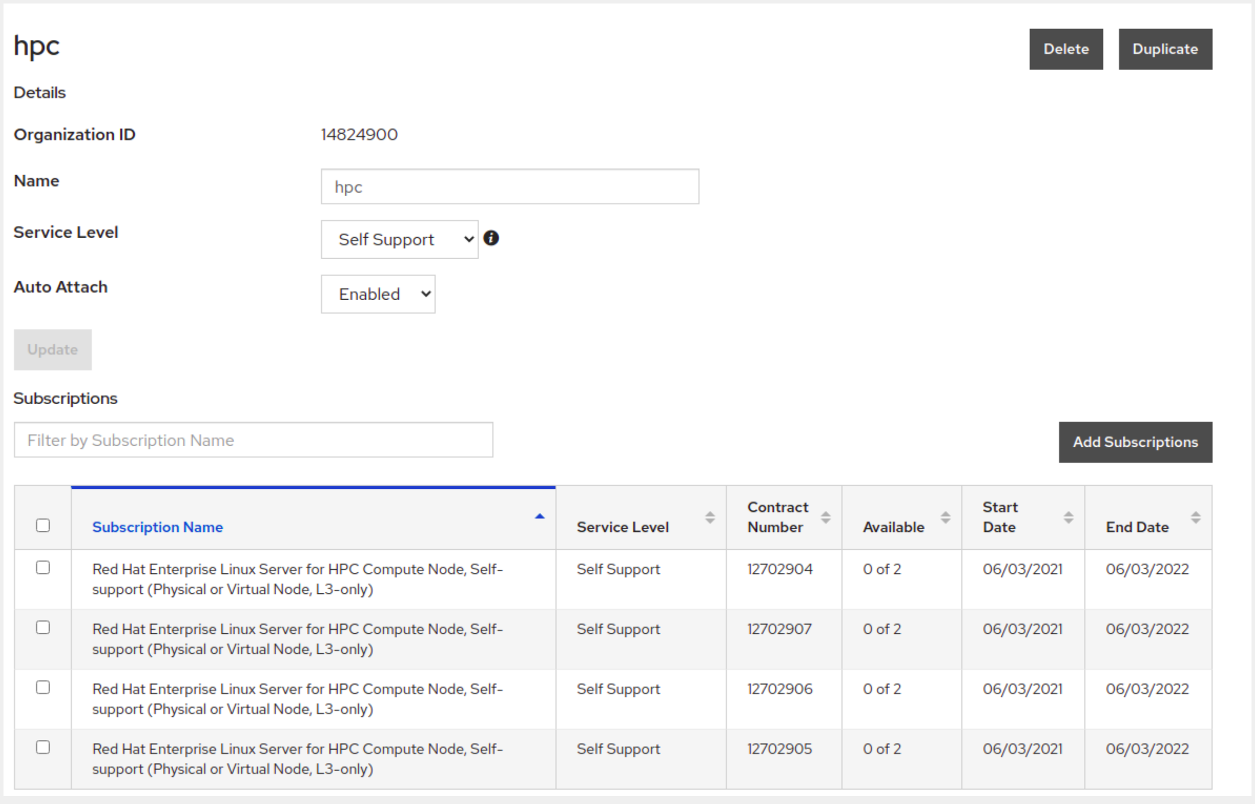
Service Level: Premium

Auto Attach: Enabled

Subscriptions

Filter by Subscription Name

□	Subscription Name	Service Level	Contract Number	Available	Start Date	End Date
□	Red Hat Enterprise Linux Server, Premium (Physical Node with up to 1 Virtual Node) (L3 Only)	Premium	12702912	1 of 1	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server, Premium (Physical Node with up to 1 Virtual Node) (L3 Only)	Premium	12702910	0 of 1	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server, Premium (Physical Node with up to 1 Virtual Node) (L3 Only)	Premium	12702914	0 of 1	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server, Premium (Physical Node with up to 1 Virtual Node) (L3 Only)	Premium	12702914	1 of 1	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server, Premium (Physical Node with up to 1 Virtual Node) (L3 Only)	Premium	12702909	1 of 1	06/03/2021	06/03/2022



hpc

Details

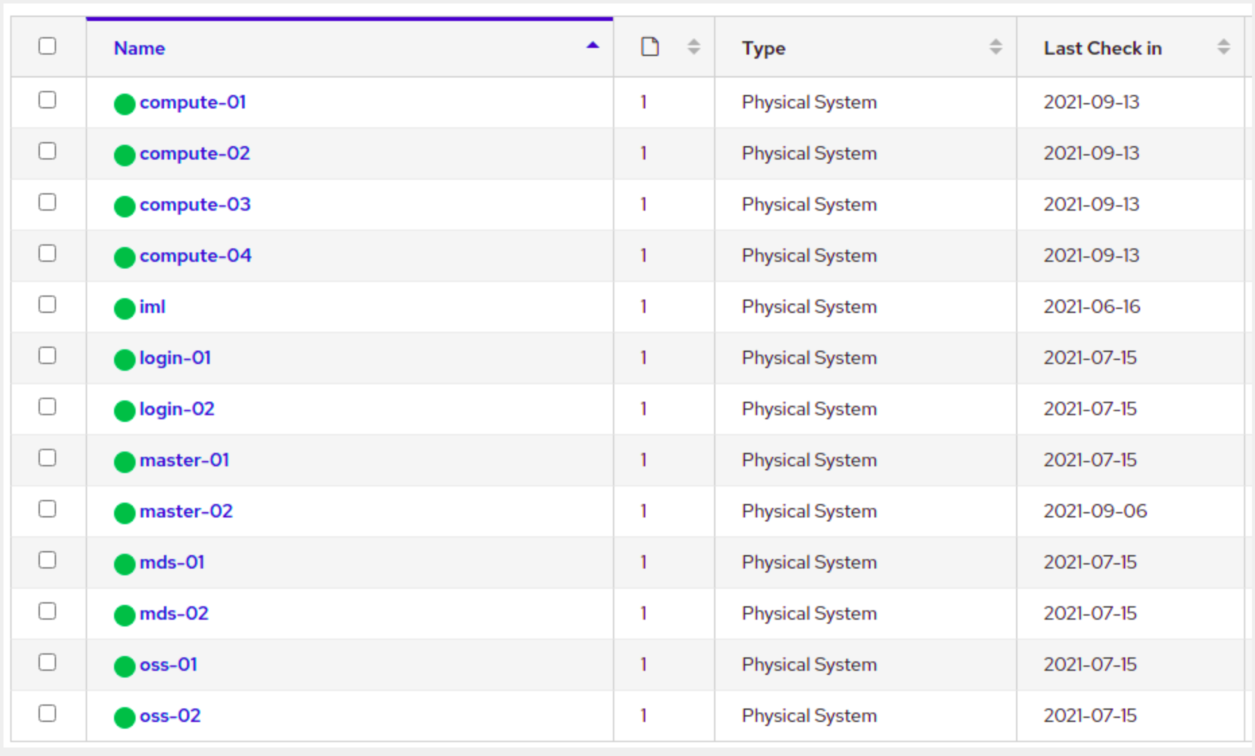
Organization ID	14824900
Name	hpc
Service Level	Self Support
Auto Attach	Enabled

Subscriptions

□	Subscription Name	Service Level	Contract Number	Available	Start Date	End Date
□	Red Hat Enterprise Linux Server for HPC Compute Node, Self-support (Physical or Virtual Node, L3-only)	Self Support	12702904	0 of 2	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server for HPC Compute Node, Self-support (Physical or Virtual Node, L3-only)	Self Support	12702907	0 of 2	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server for HPC Compute Node, Self-support (Physical or Virtual Node, L3-only)	Self Support	12702906	0 of 2	06/03/2021	06/03/2022
□	Red Hat Enterprise Linux Server for HPC Compute Node, Self-support (Physical or Virtual Node, L3-only)	Self Support	12702905	0 of 2	06/03/2021	06/03/2022

Thus to activate the subscription the following is used command respectively for servers and compute systems:

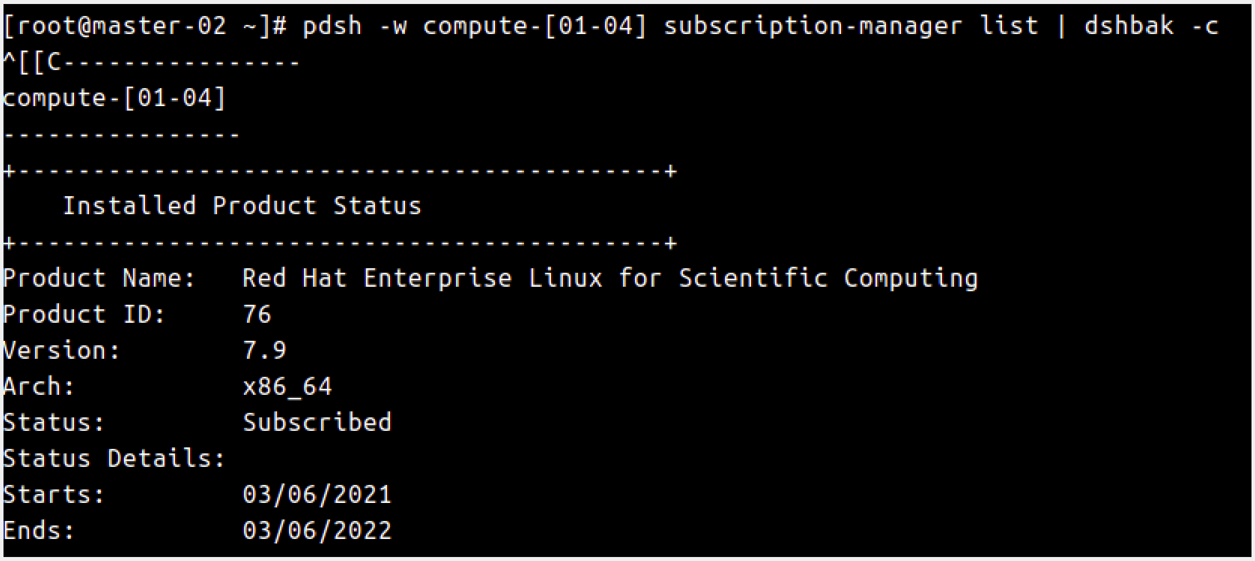
```
subscription-manager register --org=14824900 --activationkey=premium
subscription-manager register --org=14824900 --activationkey=hpc
```



□	Name ▲	🗋	Type	Last Check in
□	● compute-01	1	Physical System	2021-09-13
□	● compute-02	1	Physical System	2021-09-13
□	● compute-03	1	Physical System	2021-09-13
□	● compute-04	1	Physical System	2021-09-13
□	● iml	1	Physical System	2021-06-16
□	● login-01	1	Physical System	2021-07-15
□	● login-02	1	Physical System	2021-07-15
□	● master-01	1	Physical System	2021-07-15
□	● master-02	1	Physical System	2021-09-06
□	● mds-01	1	Physical System	2021-07-15
□	● mds-02	1	Physical System	2021-07-15
□	● oss-01	1	Physical System	2021-07-15
□	● oss-02	1	Physical System	2021-07-15

To get the subscription status from the system please execute the following command

```
subscription-manager list
```



```
[root@master-02 ~]# pdsh -w compute-[01-04] subscription-manager list | dshbak -c
^[[C----------------
compute-[01-04]
----------------
+-------------------------------------------+
    Installed Product Status
+-------------------------------------------+
Product Name:   Red Hat Enterprise Linux for Scientific Computing
Product ID:     76
Version:        7.9
Arch:           x86_64
Status:         Subscribed
Status Details:
Starts:         03/06/2021
Ends:           03/06/2022
```