NCI Cluster Documentation

Version 1.0

Table of Contents

Introduction

This document provides a step by step guide to install slurm based cluster on the NeCTAR cloud. The document assumes that the reader is already familiar with NeCTAR cloud and OpenStack.

This cluster suite is in no way a replacement of a full HPC cluster. It will only provide you with a functioning cluster. It is your responsibility to install and maintain applications on this cluster. You being root on this cluster are responsible for the security and patching the cluster.

NCI users: Slurm is not compatible with NCI's version of PBS Pro; therefore, you might have to change your job submission scripts. Slurm compatibility layer for Torque/PBS is already supplied with the image and works for standard PBS inputs.
There is no need to recompile your code as NCI cloud and Raijin share the same CPU architecture. Applications can be copied from Raijin to this cluster as long as the License permits e.g. Intel compilers are not supplied on this cluster due to License terms. You can however, copy the compiler(s) and supply your own license. You may use applications previously compiled on Raijin with Intel runtime (which is not supplied).

Disclaimer: While all care has been taken to make this cluster secure, NCI does not accept any responsibility for loss of data, unauthorised access or any other incident. It is the responsibility of the tenant to keep the cluster secure and patched.

The software is released under BSD License.

Guide for Impatient and Advanced Users

The whole process to provision a cluster is to instantiate a head-node, attach a volume to head-node, login to head-node, mount volume as /data and provision the cluster.

Following is the step by step guide for impatient and/or advanced users.

1- Launch a virtual machine based on NCI supplied image.
   a. Image name: slurmcluster-nci-{yyyymmdd). Look for latest date.
2- Attach a volume to the image from the NeCTAR dashboard. You may use the ephemeral disk supplied by certain flavours but it is not recommended.
   a. You may need to create a volume based on your storage allocation.
   b. Note the **Attached to** field. E.g /dev/vdb in this example.
3- Login to the virtual machine you launched

```
cd /root/ncicluster/
git clone https://github.com/NCI-Cloud/slurm-cluster.git
cd slurm-cluster
```

   You may skip the step of make filesystem if you already have a volume which is formatted and has data.

```
mkfs.ext4 /dev/vdb
```

   Mount the volume as /data

```
mount /dev/vdb /data
```

   Configure the head-node (one time only operation)

```
./config_headnode -i
```

   source the OpenRC file. You can download it from NeCTAR dashboard. You need NeCTAR API password.

   Launch the cluster.

```
./nci-cluster -b
```

   Start daemons and services.

```
./readyCluster -s
```

4- Start submitting jobs.
5- Read the documentation for additional functionality.

Step by Step and More Detailed Guide

The cluster provisioning consists of two basic processes. In the first process, you will launch an instance which will act as a head node of the cluster. The head node is clusters' management and login node. In the second step, you will login to the head node and provision compute nodes. We have split these processes in a number of steps.

Most of the steps in the guide that are one-time-only. For example, retrieving your API password, creating cluster volume.

Please follow the instructions carefully. For BUGS/Feature requests, you may contact help@nci.org.au.

1. Login to NeCTAR dashboard using your AAF credentials. Please note that your AAF credentials and NeCTAR passwords are two different entities. While AAF credentials will allow you to manage your virtual machines from NeCTAR dashboard, they do not let you use the command line tools/APIs.
   https://dashboard.rc.nectar.org.au

2. Select the appropriate tenant from the drop down next to NeCTAR logo.

3. If you do not have the password to access the APIs, you can follow NeCTAR documentation to download the password.

   https://support.rc.nectar.org.au/docs/authentication

Short description: If you don't know your password (or haven't set it yet) you can reset it through Dashboard under the *settings* link (top right next to signout) as shown in Figure 1.

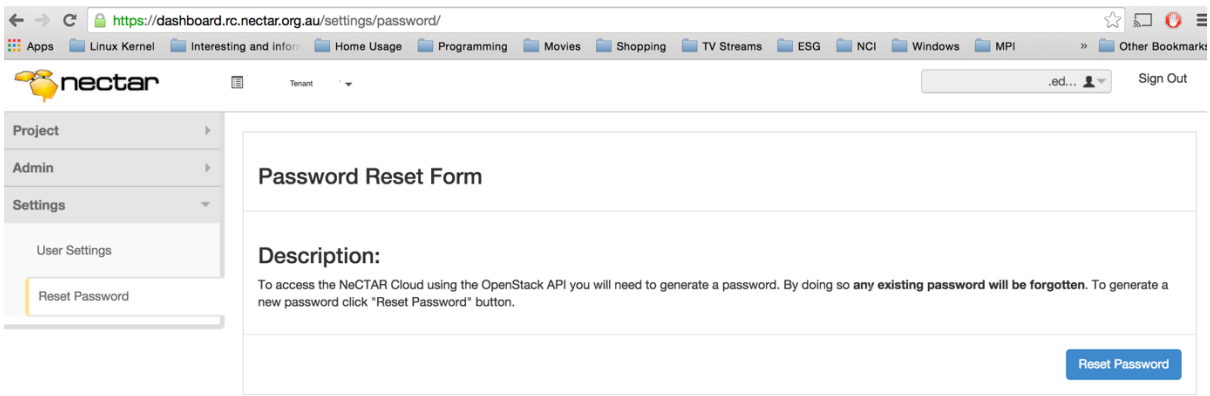

Figure 1: NeCTAR Dashboard Password Retrieval / Reset.

4. Download openrc file from the NeCTAR dashboard.
   NeCTAR Dashboard → Access and Security → API Access → Download “Openstack RC File”.

5. If you are using NeCTAR/Tenjin for the first time, Click “Access & Security” and create/add your KeyPair. On Linux or Mac, you can “cat” your public key and paste it here. You may give it any name.
   e.g `cat ~/.ssh/id_rsa.pub`

   This is one time only operation. However, depending on your workflow/security model you may choose to have a number of key pairs.

6. **Optional:** Install OpenStack tools on your workstation (or virtual machine). [*You **may skip** this step and use NeCTAR dashboard as described in step 7 (dashboard is recommended for new users)]*

   a. One Time Only for Mac Users. This should work on Linux. Windows users please proceed to Step 5 or install python and related packages yourself.
      - `sudo easy_install pip`
      - `sudo pip install virtualenv`
      - `cd my_project_folder`
      - `virtualenv venv`
      - `source venv/bin/activate`
      - `pip install python-novaclient python-keystoneclient python-cinderclient`
      - source myNeCTAR-project-openrc.sh (Your Openstack RC File)
        `source myproject-openrc.sh`
      - Give your API Password from Step 2
      - Test if everything works "`nova flavor-list`"
      - Add your Key to OpenStack.
        - nova keypair-add --pub-key ~/.ssh/id_rsa.pub mylogin key
        - nova keypair-list (Your key should be in the list)

   b. Launch the HeadNode using your Key
      - Check which image to use ; use the one with latest date
        `nova image-list | grep nci`
        You expect something similar:
        `…826415db-d9d7-4e6b-b84f-7656501f2537 | slurmcluster-nci-20150923 ..`
      - Launch the image with appropriate flavor. Note that the name of the login node should be unique to your project. i.e. never launch two nodes with the same name. Please check your availability zone. If not sure, try NCI. ☺
        `nova boot --image slurmcluster-nci-20150923 --flavor m2.medium --key_name myloginkey --availability_zone NCI clusterhead`
      - Wait for a while and then ssh into your virtual machine
        `nova list` (check the IP address of the virtual machine)
        `ssh root@XX.XX.XX.XX`
      - You should be logged into the virtual machine.

   You may skip the next steps and continue to Step 8

7. Launch a virtual machine using NeCTAR dashboard. This virtual machine will act as your login node, management server and the NFS server for the cluster.

   a. NeCTAR dashboard->Instances-> "Launch Instance"
   b. Image: "SlurmCluster-NCI-(with date appended). Latest date is recommended if you find multiple images. E.g. slurmcluster-nci-20150923

   If you wish to use your own image, please refer to "Setup my own image" section of the documentation. You should choose Centos 6.X flavour from the official

NeCTAR images as we only support RHEL 6 derivatives. Debian based distributions are not supported.

c. Choose appropriate flavour. E.g. m2.medium for a small cluster.

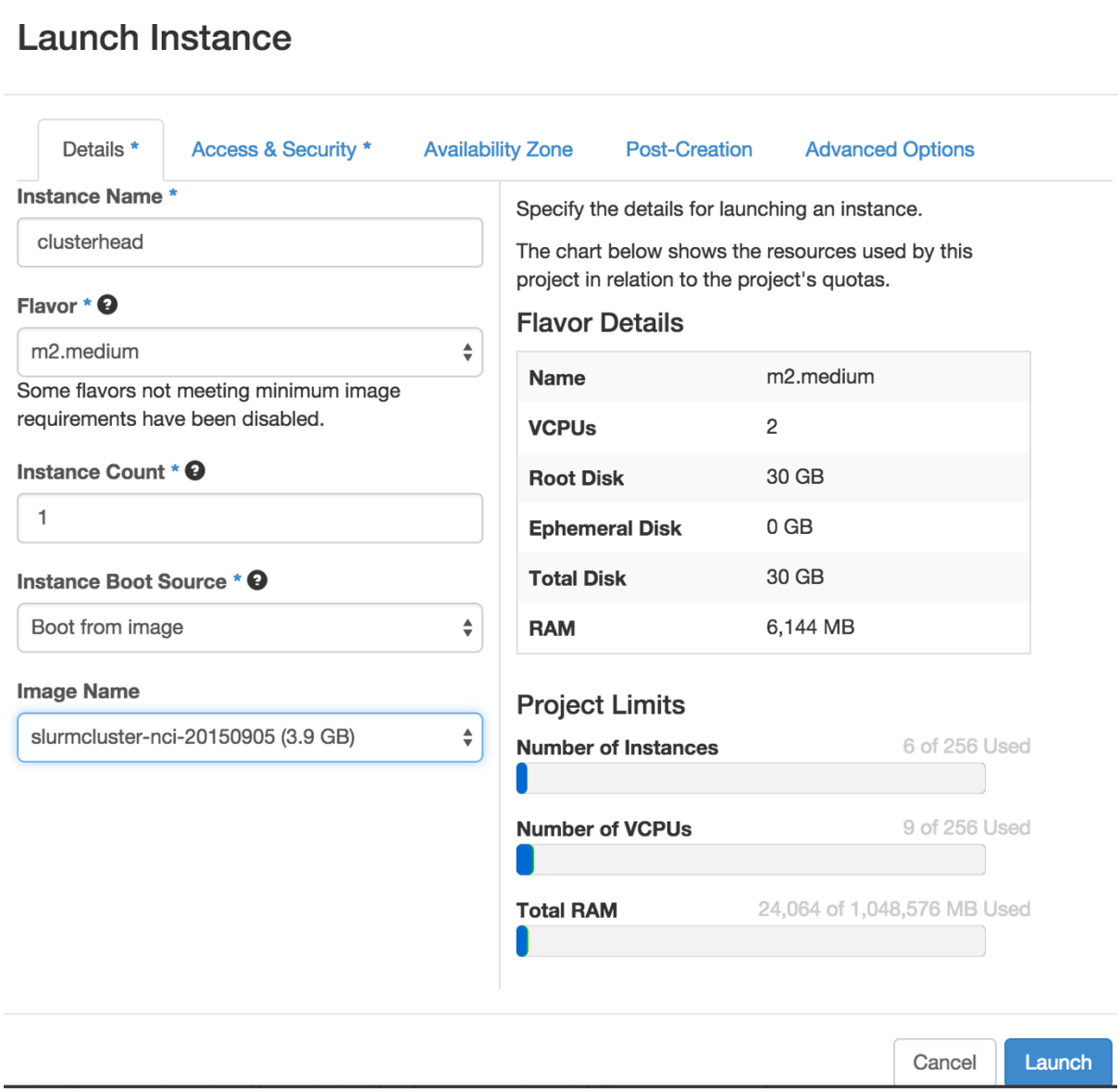

Figure 2(a): Launch Instance

d. In "Access & Security" tab, select the public key you later wish to use to login to the instance.
e. Make sure that you launch the instance at the appropriate availability zone. E.g NCI. This is shown in Figure 2(b).
f. Once you launch an instance, please wait and the dashboard will refresh to give you the IP address of the instance.

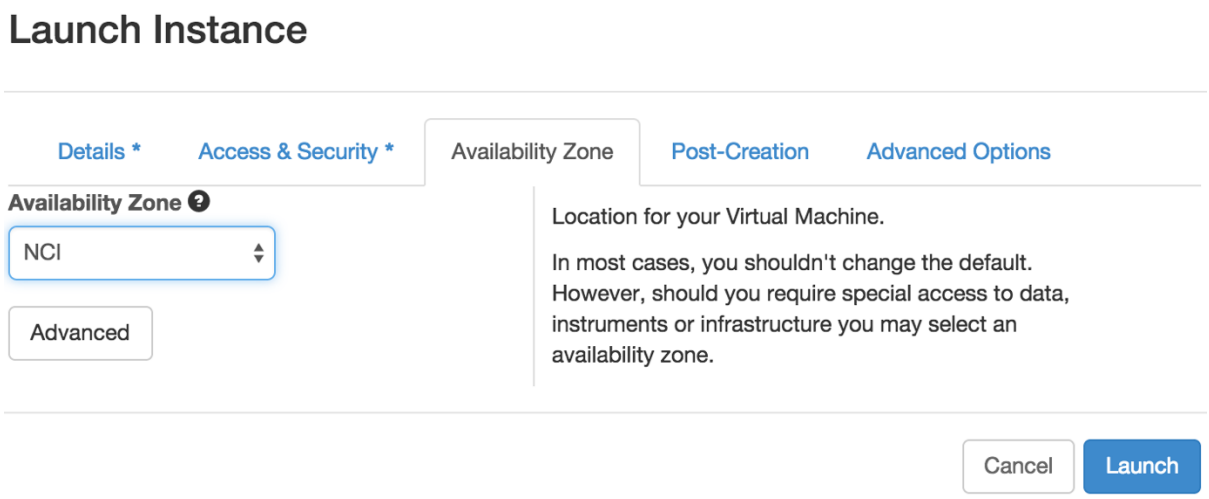

Figure 2(b): Select Availability Zone

8. Login to the virtual machine using the ssh key. This is to make sure your instance has come up without any problems.

```
ssh root@ip.address
```

9. Once you login to the virtual machine, it should have a root filesystem (/dev/vda) and possibly an ephemeral disk (/dev/vdb). On NCI private cluster, you will always get an ephemeral disk but few NeCTAR flavours do not have an ephemeral disk. We highly recommend that you use a persistent volume as an NFS volume. You need to ask for volume storage at the time of new tenant request.
    a. Create a volume at NeCTAR dashboard→ Volumes→"Create Volume". *Ensure that the volume is on the same availability zone as your virtual machine otherwise your data access will be seriously impacted.*
    If you already have a volume, you can skip this step.

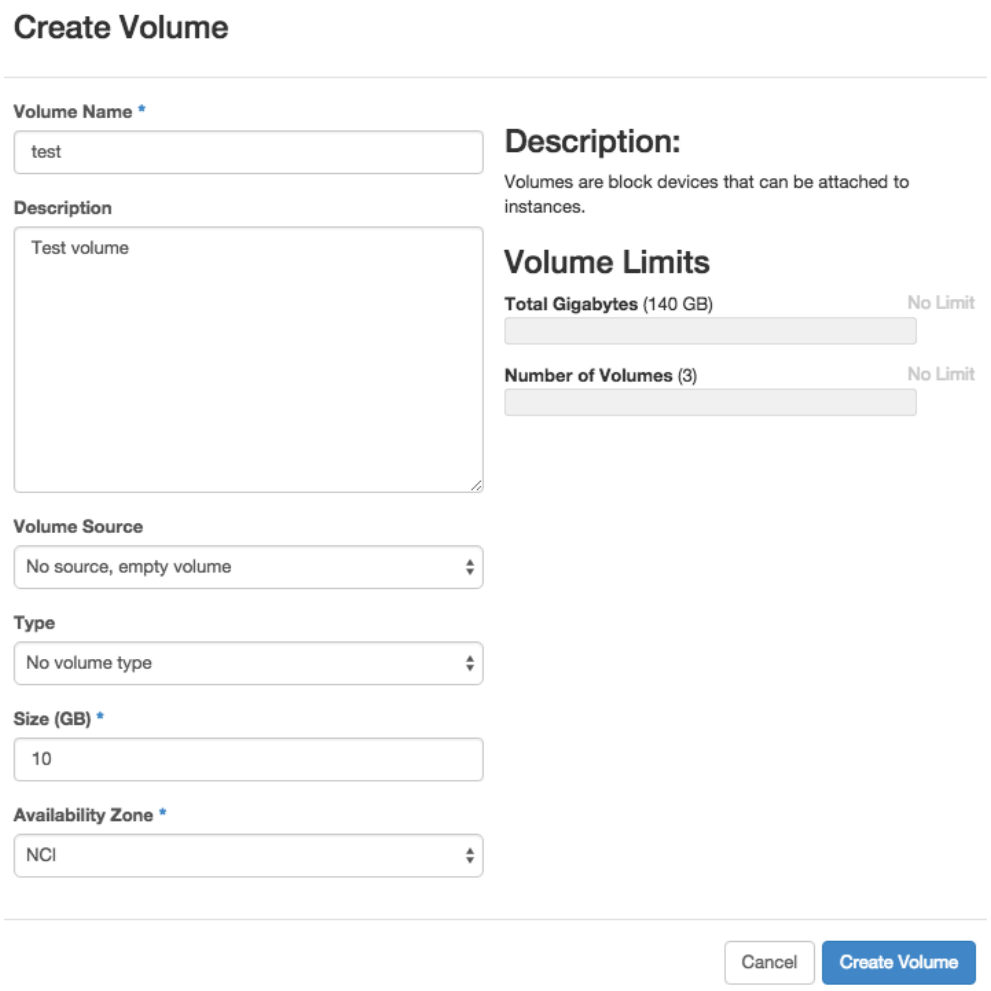

Figure 3 (a): Create a Volume.

b. Attach the volume to the virtual machine as shown in Figure. ("Edit Volume" Drop-down-> "Edit Attachments". The volume should appear as /dev/vdb (or /dev/vdc if you have ephemeral storage) on your virtual machine.

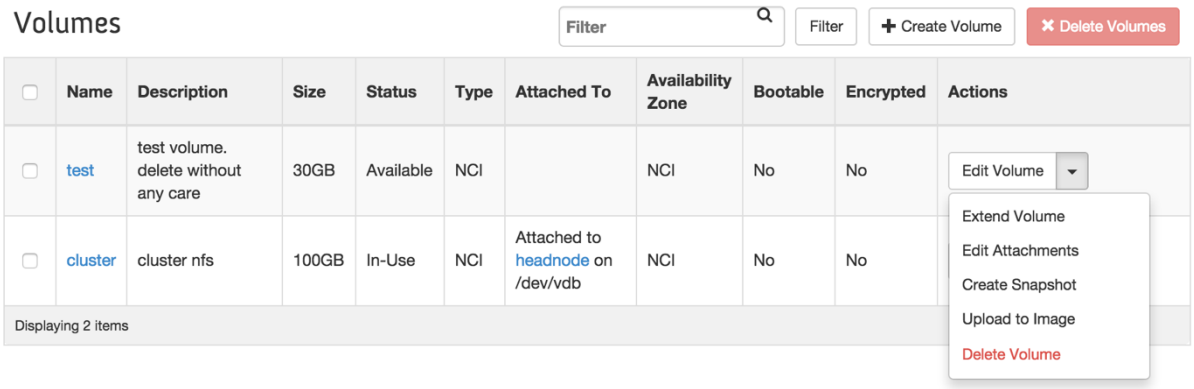

Figure 3(b): Attach volume

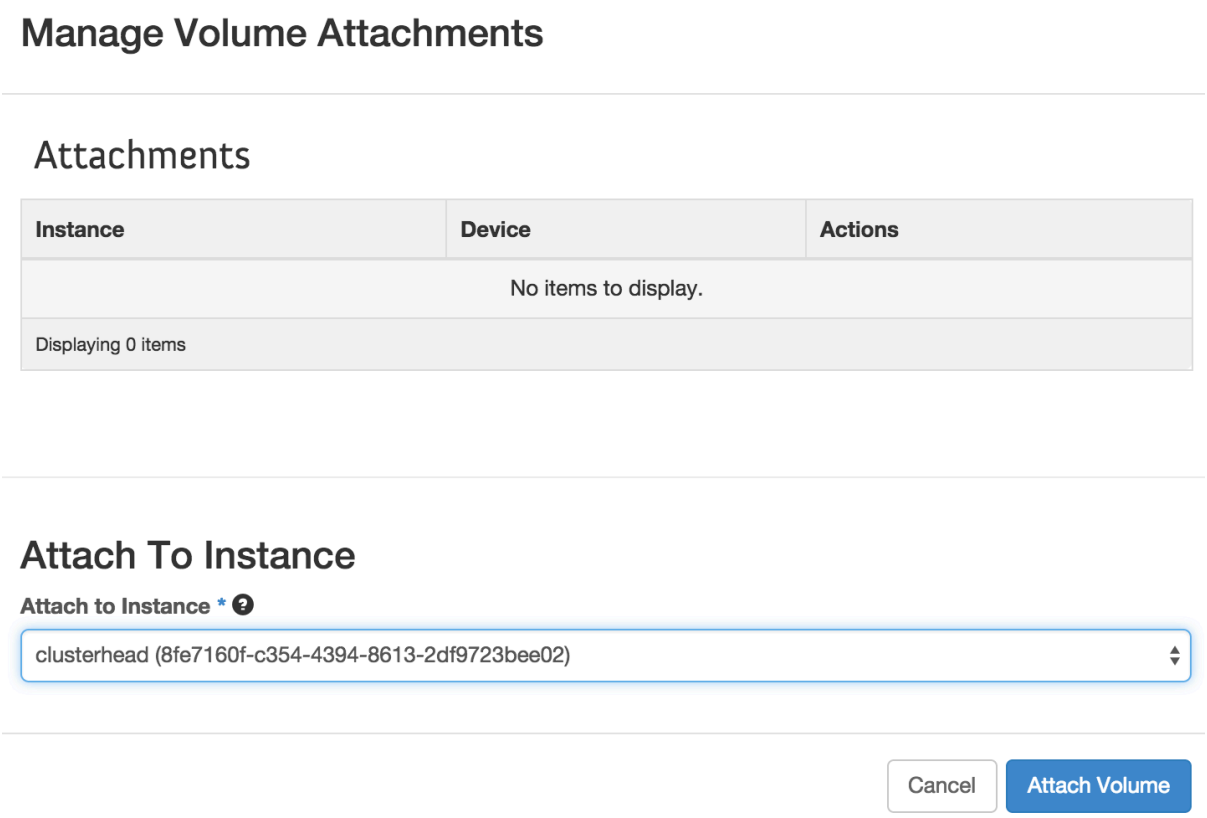

Figure 3 (c): Attach volume to your head node.

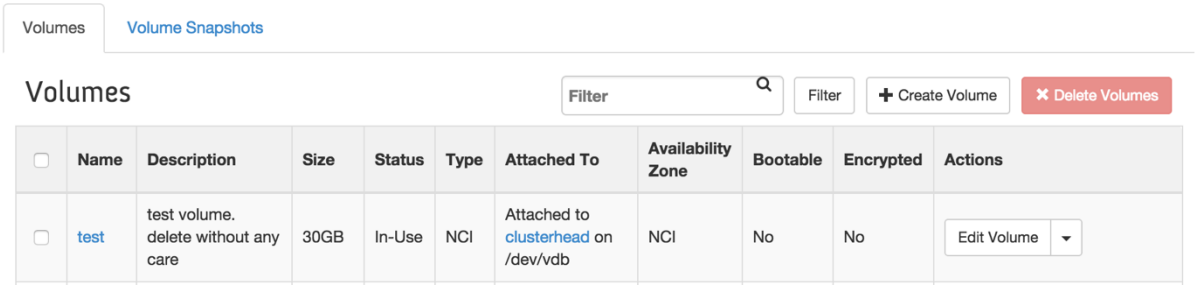

Figure 3 (d): Dashboard providing details of a volume attached to an instance as /dev/vdb.

10. SSH to the head-node and create a file-system on the volume your just attached. You may use `fdisk -l` to determine the name of the volume. For example, in Step 7, we created a 10GB Volume. Running fdisk -l suggests that the volume /dev/vdb is the volume which has ~10 GB size.

    Example is shown below

```
[root@headnode ncicluster]# fdisk -l

Disk /dev/vda: 32.2 GB, 32212254720 bytes
22 heads, 16 sectors/track, 178734 cylinders
Units = cylinders of 352 * 512 = 180224 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00066a00

   Device Boot      Start         End      Blocks   Id  System
```

```
/dev/vda1   *           6      178734    31456160   83  Linux

Disk /dev/vdb: 10.7 GB, 10737418240 bytes
16 heads, 63 sectors/track, 20805 cylinders
Units = cylinders of 1008 * 512 = 516096 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x00000000
```

11. Create file-system on /dev/vdb as follows:

```
[root@awesome /]# mkfs.ext4 /dev/vdb
 mke2fs 1.41.12 (17-May-2010)
```

Note: Be very careful of the name. /dev/vdb or /dev/vdc.

Also, if you already have an existing volume and you do not want to destroy data, please skip this step and simply mount the volume as described in the next step.

12. Mount the volume on your virtual machine. *You cannot launch cluster without /data mounted*.

```
[root@awesome /]# mkdir -p /data; mount /dev/vdb /data
```

NOTE: You 'must' mount the volume on /data.

13. We recommend that you make this entry persistent in /etc/fstab. Edit /etc/fstab and enter the following text.

```
/dev/vdb     /data  ext4   defaults     0 1
```

14. Download Slurm-Cluster tool from GitHub to /root/ncicluster folder.
    https://github.com/NCI-Cloud/slurm-cluster
    Or simply clone the repository:

```
cd /root/ncicluster
git clone https://github.com/NCI-Cloud/slurm-cluster.git
```

15. Change directory to your cloned repository. Execute "./config_headnode.sh" script. This script has two modes.
    a. -i [--from-image]: Recommended if you are using pre-built image from NeCTAR/Tenjin e.g. slurmcluster-nci-20150923
       - Configures NFS – Note that it requires /data to be mounted.
       - Copy prebuilt Slurm and OpenMPI to /opt/slurm and /apps/openmpi/1.10.0 respectively. Prebuild packages are already present at in /contrib folder of the image.
       - Generate Keys (/root/id_rsa and /etc/munge/munge.key)
       - Perform yum update –y
    b. -b [--build-image]: Use this option when building your own image. e.g. you are building an image from official NeCTAR images
       - Install packages.
       - Install OpenStack command line tools.
       - Configure NFS server
       - Download, compile and configure Slurm.

- Download, compile and configure OpenMPI (version 1.10.0).
- Copy compiled packages into /contrib.
- Generate Keys
- Perform yum update -y

16. Source OpenRC file and give your API password.

```
source myproject_openrc.sh
```

When prompted, give your API Password.
To ensure that you have given the right password, please type the following command on the command line

```
nova flavor-list
```

You should see output with flavor ID, Name, Memory MB etc.

Tip: Download the file on your desktop and copy it to your head-node.

17. Edit cluster.cfg file supplied with the installation.

This is the most critical step and might require you to perform 'nova' command line or extensive usage of OpenStack dashboard.

The configuration is ready for NeCTAR. You only need to change the AVAILABILITY_ZONE (default is NCI) or optionally change CLUSTER_COMPUTE_SIZE if you want more than two compute nodes for your cluster.

Sample configuration file is shown below. Please note that we have already provided you with the cluster.cfg file.

```
[Cluster]
#Name your cluster. Any unique name is fine. Try not to use spaces or special characters
CLUSTER_NAME=mycluster
#all lower case – compute node name prefix
INSTANCE_NAME=nodex
#Start of the compute node name e.g. nodex1
CLUSTER_RANGE_START=1
#Cluster size to launch. 2 means that two compute nodes will be launched. First node will be named
nodex1 and second nodex2.
CLUSTER_COMPUTE_SIZE=2
# Max size for the cluster.
MAX_CLUSTER_SIZE=10
# Your NeCTAR/Tenjin/OpenStack Key used for launching instances.
# You can get the name with 'nova keypair-list' command. The cluster tool will create your key
automatically if it does not exists. Currently we only use the default keys (private
/root/.ssh/id_rsa and public /root/.ssh/id_rsa.pub).
KEY_NAME=mykey
# Image name on the Openstack to use as a base
# NCI will provide you with the image name. You can create your own image based on RHEL. Section
Create your own image later in the document.
# Use nova image-list to determine the flavor list.
IMAGE_NAME=slurmcluster-nci-20150906
# Flavour name. "use nova flavor-list" to see what flavors are available.
# For NeCTAR users: https://support.rc.nectar.org.au/docs/launching-instance
FLAVOUR_NAME=m2.medium
# Availability zone; on NeCTAR e.g. is NCI, pawsey-01 etc
AVAILABILITY_ZONE=NCI
# Your Tenant network. On NeCTAR if you are using NCI just use NCI. This value is not used. This
may change once NeCTAR cloud moves to Neutron.
```

```
# If you are using Tenjin@NCI, then "nova network-list" command will give you the network to use.
It is usually with your Tenant name.
NETWORK_NAME=nectar
# Do not change the security group name. It will remain X for ever. :)
SECURITY_GROUP=X
# Not used in NeCTAR. You need to manually create and attach the volume to the headnode.
NCI_CLUSTER_VOLUME=CLUSTER_VOL
# Deprecated.
COMPUTE_CLOUD_INIT=cloud_init.sh
# Slurm installation; no need to change
SLURM_HOME=/opt/slurm/
# Cluster opt folder; no need to change
CLUSTER_OPT=/opt/
# Cluster NFS share. Do not edit as it is fixed to /data
CLUSTER_NFS=/data
#Your slurm partition. No need to change.
SLURM_PART=Cloud1
```

18. You may now provision the compute nodes.
    `./ncicluster.py -b`
    You will see output detailing the progress.

19. You may need to wait for a while for the compute nodes to build. You may use `nova list` to see the state of the built process. Once your cluster nodes are up, you need to start services. Use the following script.
    `./readyCluster.py -s`

20. You are now ready to submit jobs. We recommend using slurm user if the cluster is only for you. You can create individual users etc. However, this is not in the scope of this document.
    a. `sinfo` on headnode should give you status of the nodes.

    ```
    PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
    Cloud1*      up   infinite      2   idle node[1-2]
    ```

    b. `squeue` gives status of the queues.
    c. `module avail` to list available application environment modules.
    d. `Module load openmpi` will load OpenMPI module.
    e. `sbatch` will submit a job. E.g `sbatch mpi.job`

21. We have build clustershell into the image. This is replacement for pdsh.
    E.g. `clush -a service iptables status`

Submit your first job

To ensure our cluster is capable of running MPI jobs, try the following (copy and paste friendly)

```
su -  slurm
cd /short/slurm
wget http://mvapich.cse.ohio-state.edu/download/mvapich/osu-micro-benchmarks-5.0.tar.gz
module load openmpi
tar -xvf osu-micro-benchmarks-5.0.tar.gz
cd osu-micro-benchmarks-5.0
./configure && make
cd /short/slurm
mkdir -p /short/slurm/jobs
cd jobs
cat > mpi.job << EOF
```

```
#!/bin/bash -l
#
# Request 5 hours run time
#SBATCH -t 5:0:0
#
#SBATCH --nodes 2
#SBATCH -n 2
#
module load openmpi/1.10.0
mpirun /short/slurm/osu-micro-benchmarks-5.0/mpi/pt2pt/osu_latency
EOF
sbatch mpi.job
squeue
# Now check the output file in your working directory.
```

Cluster Maintenance

Note: Please perform regular updates on the cluster. We highly recommend updating the compute nodes immediately after provisioning. You should also frequently patch the system.

To update the headnode: `yum update -y`
To update the compute cluster: `clush -a yum update -y`

Shutdown Cluster Nodes

1. To shutdown compute nodes of the cluster, please issue with following command.
   `./nci-cluster.py -d`
   This will effectively terminate compute nodes, remove NFS exports and iptables entry.

Extend or Shrink the Cluster

1- To extend the cluster, please issue the following command
   `./nci-cluster.py -e 2`
   This will add two more nodes to the cluster. You need to restart services on the newly provisioned nodes e.g. `./readyCluster.py –n node10`

2- To shrink the cluster, please issue the following command
   `./nci-cluster.py -s 2`
   This will terminate last two instances. Please make sure that you have drained the nodes in slurm so that there is no loss of jobs.

User and Group Management.

Currently, the cluster tool gives you basic user and group administration. Create a user or group the standard linux way on the headnode. Use the script, `syncCluster.py` to sync to the cluster. LDAP based cluster might be released if there is a demand.

The login is kept with ssh keys. The administrator may change these settings. Currently, the administrator might need to add user keys in users $HOME/.ssh/authorized_keys.

Create image based on NCI supplied image

Launch an instance based on NCI supplied image. Add or remove the packages, reconfigure configuration files as required and use `create_image.sh` script to prepare instance for snapshot. Shutdown the instance and create a snapshot.

Built Image from scratch

We recommend that you use NCI supplied image. However, in case you need to build your image, please follow these guide lines. You might need to debug your image. This section is only provided to encourage open development.

1- Launch a virtual machine based on official NeCTAR Centos 6.X image.
2- Login to the virtual machine and download cluster tools (See above section)
3- Install packages you want.
   Example:

```
yum groupinstall 'Development Tools' –y
yum install munge munge-devel hwloc hwloc-devel environment-modules
```

4- `./configure_headnode.sh -b`
5- Copy openmpi build, slurm build and slurm-home to contrib folder.

```
mkdir -p /contrib
rsync -av /opt/slurm /contrib
rsync -av /apps/openmpi /contrib
rsync -av /home/slurm /contrib/slurm-home/
rsync -av /apps/Modules /contrib/
```

   Create `/etc/init.d/slurm`
   Create `/etc/profile.d/slurm.sh` (or create Module)
6- Take an image snapshop
   a. `./create_image.sh`
   b. Go to NeCTAR Dashboard, shutdown the instance and create a Snapshot.
7- You can now launch the Snapshot (or Image as there is no distinction in OpenStack) as your head node or the compute node.

Possible features in the next version

- Ldap in the headnode so that user management is made easy. (depends on user interest)
- NCI PBS Pro compatibility mode. We already supply basic compatibility with Torque.
- MySQL based accounting. Currently it is disabled but cluster admin can enable it.
- Asynchronous compute node provisioning.

Known Bugs

- Multiple entries for headnode IP address in the firewall if someone continues to run ./config_headnode.sh script. (Not dangerous but still a bug; and you are not meant to configure headnode multiple times)

Slurm Admin Cheat Sheet

Basic commands that are required in maintaining the cluster. Please refer to slurm documentation for details.

1- Bring node from down state to idle state.

```
scontrol update NodeName=node6 State=RESUME
```

2- Drain a compute node. Un-drain with state IDLE or RESUME as shown in Step 1.

```
scontrol update NodeName=node6 State=DRAIN Reason="MCE- Memory"
```