UPPSALA
UNIVERSITET

v1.0, June 22, 2016
Author: Jerker Nyberg von Below

SOP – Linux HPC Cluster Disaster Recovery

1 What is this?

This is intended to be both a description on how the cluster is set up and a disaster recovery document. The idea is that this is a general document without any secrets. The specific details of the cluster is stored in another document.

2 General description of system setup

2.1 File server

The cluster is totally depending on the file server. Without the file server nothing works. The file server is running the following services:

- DHCP – to give all execution nodes network information. This references the TFTP-server
- TFTP – to be able to boot the execution nodes over network to start a kickstart installation
- NFS – This is the way all data on the cluster is shared to the nodes. This traffic is neither authenticated nor encrypted. It is only protected by being only available on a non-routed local network
- NAT-router – The file server also is running as a NAT-router for the execution nodes between the cluster network and the outside network
- SSH – The SSH-server on this node is only possible to login for the system administrator.
- MySQL – Slurm store data here.

2.2 Cluster nodes

The cluster nodes are running the following services:

- Winbind – The execution nodes are connected to a shared user database in the university Active Directory called USER-AD
- SSH – It is possible to login at all the execution nodes for all users. Login is limited to a certain group in the USER-AD
- Slurm – The nodes accept jobs

2.3 Execution nodes

The execution nodes run nothing else than the above.

2.4 Test node

The test node does not accept job via Slurm

2.5 The head nodes

The head nodes are a sub-category of the execution nodes with the following exceptions:

- They have a second physical network interface connected to the outside network. This network also have the default gateway.
- This means they are available for logging in via SSH and SCP.

2.6 Two slurm master nodes.

The Slurm master nodes are a sub-category of the execution nodes. They are run in two different chassis so they are not in the same physical hardware for availability reasons.

- They run the Slurm scheduler.

3 Network setup

These three networks are involved:

- Outside network
  The file server and head nodes are connected to the outside network.
- Cluster network
  The file server and all execution nodes are connected to the cluster network.
- Console network
  All nodes with a separate console network port (IPMI) are connected to the console network. This network is not routed. (It might be mentioned that the current machines do not very well at all using the network consoles, most probably due to time limitations when setting the machines up. This is not active.)

4 Physical infrastructure

The cluster is located in the computer room at BMC. The computer room has dual battery backed up UPS connected to a diesel backup generator. All nodes are connected to two sources of power.
The network is connected to a single outside switch (which is connected to the BMC-router) and a single cluster network switch. Both switches and the fiber are single-point-of-failures. However it is not very often the switches or the router breaks down. Estimated failure rate is in the 100 year or more period (none of the UU main campus routers have failed the last 20 years, but one was stolen once AFAIK). The university backbone routers which the campuses are connected too has never so far failed but there have as far as I know been a few times with missconfigurations.

5 Backup

The cluster is backed up via these two possible ways;

5.1 TSM (Tivoli Storage Manager)
This is the shared university backup system, managed by IT-division. This is a secure system storing the data for at least 10 years, fulfilling all possible government regulations. The only drawback is that is not dirt cheap (but it is reasonably priced) and that restore of many small files are very slow.

5.2 ZFS with Rsync
This is separate dirt cheap backup system, just synchronizing the content of the cluster file server to a separate physical PC running ZFS. The data is stored using snapshots. The idea is that this backup is faster but not as reliable as TSM.

6 Installation of cluster nodes

6.1 The file server
The file server is manually installed

6.2 All execution nodes are kickstart-installed from the file server.

During the last step of the kickstart-installation the nodes run a postrun script. This script is what in practice does all the configuration of the clustger nodes. This postrun script can, and is run, every time a configuration change is done to the cluster nodes. This way one is assured that all the cluster nodes have the same state in their configuration

7 Exceptions

7.1 There are certain configuration settings that has been done manually to the cluster nodes. The main head node is running the ThinLinc graphical thin client system. This means it is also running a desktop environment.

8 What happens when things break down

8.1 If the file server breaks down the main focus is the file systems it is running. The integrity of the file systems are of most importance. When the file systems have been restored, on the current file server or any other, the cluster nodes can mount the file systems again and start to run

8.2 If the cluster nodes break down the general idea is that the cluster can continue to run even when single nodes are not responding. If all head nodes are gone, then of course users can not log in. If both Slurm master nodes are gone all batch processing via Slurm stop working.

9 External dependencies

- University Active Directory USER-AD
- University DNS resolvers
- Network via switch via BMC-router via UpUnet via SUNET
- Power
- Cooling

10 How to restore the cluster to a running state after different kinds of failures and problems

- Cluster restart following complete power failure
- Single node hard drive (SSD) failure
- File server single hard drive (SSD) failure
- File server RAID set failure
- Single node other hardware failure
- File server very slow high latency
- Cannot login to cluster due to name server lookup or Active Directory problems
- Network not accessible

11 Who are responsible for what?

- Server room cooling and power – Akademiska Hus
- Server room in general – UUIT/BMC-Hall/BMC-IT
- Network BMC-router – UUIT/Netsupport

- Network server room – UUIT/Netsupport/BMC-IT

12 Design thoughts

Keep the file server simple. Normal PC file server, nothing special. Do not share the file server but keep it dedicated to the cluster. This reduces service periods.

Have both fast disk based restore available and slower but reliable tape backup via external service. The fast disk based backup handle most cases but the reliable tape based backup catches what is left.