Cluster Development - Rocks V

After receiving the new cluster hardware, we decided to upgrade to Rocks 5.0 and the latest version of Condor, Geant4, etc. This guide will outline the steps installing the essential OS and software.

Contents
1. Rocks V Install
2. NAS Setup
3. Condor Installation
4. Kickstart Modifications
5. Geant4 Installation
6. Network Tweaks
7. Monitoring Systems

Rocks V Install

The rolls used in the new install are Area51, Base, Ganglia, HPC, Java, Kernel, OS (1&2), Web-server. We used the standard compute cluster Jumbo DVD.
Insert the DVD and type frontend at the prompt.

Port number PP4-P33 in the High Bay has been configured to assign the correct IP address to the MAC address of the frontend, so set Rocks to DHCP and accept the IP information. Verify that the IP is 163.118.206.157.
The following information is essential:
Hostname: uscms1.fltech-grid3.fit.edu
Cluster: FLTECH
Certificate Org: FLTECH
Melbourne
Florida
US
Location: N28.10 W80.63

Due to issues running out of space in the root partition of the frontend, we manually partitioned the drive with the following:
/   100GB
/var   10GB
swap   16GB
/export   <Remainder>

The install should complete automatically.

NAS Setup

The new NAS is a 16-drive hardware RAID machine that we configured as RAID6 for a total size of 9.6TB. Boot from Knoppix and configure a single partition with no filesystem. Also make sure that the 30GB carved boot sector is set to the msdos drive label, and the large partition to the gpt drive label.
There are few filesystems that can support such a large volume, so we decided to go with XFS. This required recompiling the kernel. However, some sources indicate that the 2.6.18 kernel, that is standard in Rocks V, can be subject to kernel panics with large volumes. We decided to upgrade to the latest kernel 2.6.25 on all NAS devices.

First install Rocks on the smaller boot partition, and ignore the large partition. On the NAS you only have to create a root (/) partition and swap file.

Once Rocks is installed and booted, get the latest kernel using wget (from kernel.org) and untar it inside:

```
# /usr/src/redhat/BUILD
```

Copy the old configuration file over:

```
# cp /usr/src/linux-2.6/.config /usr/src/redhat/BUILD/<kernel name>/<kernel
name>
```

Now make sure you have the following RPM packages installed:
unifdef, ncurses-devel, xfsprogs

Run the following commands as root:

```
# make clean
# make oldconfig (select all defaults)
# make menuconfig ---> You must activate all XFS tabs under File systems.
Make sure they are starred.
# make bzImage
# make modules
# make modules_install
# make install
```

Reboot and select the new kernel to make sure it works.
Now make the new kernel default by editing the default line in /boot/grub/grub.conf
Run the following command to build the large partition as XFS:

```
# mkfs.xfs /dev/sdb1 -f
# mount /dev/sdb1 /nas0
```

Now set up NFS:

On the NAS, edit /etc/exports with the following line

---> `/nas1 10.0.0.0/255.0.0.0(rw,no_root_squash,sync)`

Edit /etc/fstab with the following line

---> `/dev/sdb1    /nas0           xfs           defaults  1 2`

On the Frontend, edit /etc/fstab with the following line

---> `nas-0-0.local:/nas0   /mnt/nas0      nfs  defaults  0 0`

The large XFS partition should now be accessible by the Frontend (after reboot).

Condor Installation

We created a script to completely automate the installation and configuration of Condor on the Frontend. A copy is attached, and is also stored on the NAS.
Create a directory in root's home and place the Condor RPM and setup script inside.
Run:

`# sh setupCondorforBoot`

Check the configuration:

`# /etc/init.d/condor start`
`# ps ax | grep condor`

The daemons should all be running (master, negotiator, collector, startd, schedd)
If you want, you can disable the Frontend as a compute node with the following command:

`# condor_configure --type=manager,submit`

Kickstart Modifications

#Before running rocks-dist dist, be sure that /sbin and /usr/sbin are in the PATH#
The node installation is completely automated by the extend-compute.xml and links-compute.xml script, as well as replace-partition.xml for compute node partitioning. A copy of these are attached and stored on the NAS.
You must place the condor RPM inside /export/home/install/contrib/RPMS and the extend-compute.xml and replace-partition.xml inside /export/home/install/site-profiles/5.0/nodes. Also place links-compute.xml inside /export/home/install/site-profiles/5.0/graphs/default.
Run:

`# cd /home/install`
`# rocks-dist dist`

Ensure that stage2.img, product.img, and updates.img are inside /home/install/rocks-dist/lan/x86_64/images. Also ensure that the kickstart graph is correct (connecting X11 and devel to compute appliance).

Install a node with the DVD or kernel boot CD and run:

`# ps ax | grep condor`

Ensure that startd, master, and schedd are running. The node's CPUs should show up on condor_status as Unclaimed.

Geant4 Installation

Create a geant4 user with home directory on the new NAS. Untar Geant4 into an install directory inside. Also add the openmotif package. If you want to use CRY you should install it before Geant4
Run:

```
# ./Configure --build
```

The installation is interactive, so you must select the libraries that you will need. We chose the following:
Build shared libraries and global libraries only.
Build XAW
Build XM_Motif
Use Dawn Driver
Build OpenGLX and OpenGLXM
Build Raytracer
Build VRML
Build zlib
Build Analysis

Now run ./Configure
Geant4 should be installed. You should install Root as well, but it is very easy, just follow the README.
Add the following lines to the geant4 user's .bash_profile:

```
source /mnt/nas0/home/geant4/geant4Install/geant4.9.1.p02/env.sh
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/mnt/nas0/home/geant4/
geant4Install/geant4.9.1.p02/lib/Linux-g++:/mnt/nas0/home/
geant4/geant4Install/bin/lib
```

In order for the jobs to run on the nodes, copy the libXm.so and libXp.so files from /usr/lib64 to /mnt/nas0/home/geant4/geant4Install/bin/lib

Network Tweaks

As of November 2008, we have upgraded our link to Gigabit. The following outlines TCP and other networking tweaks performed.

As instructed by UF, the following files were modified:

```
echo 16777216 > /proc/sys/net/core/wmem_max
echo 16777216 > /proc/sys/net/core/rmem_max
echo "4096000 8738000 16777216" > /proc/sys/net/ipv4/tcp_rmem
echo "4096000 8738000 16777216" > /proc/sys/net/ipv4/tcp_wmem
echo "4096000 4096000 4096000" > /proc/sys/net/ipv4/tcp_mem
```

The following are tweaks based on BDP Calculations.

```
echo "1024000 2048000 4096000" > /proc/sys/net/ipv4/tcp_rmem
echo "1024000 2048000 4096000" > /proc/sys/net/ipv4/tcp_wmem
echo 4096000 > /proc/sys/net/core/wmem_max
echo 4096000 > /proc/sys/net/core/rmem_max
```

The BDP calculated tweaks appear to work well, however in the 2.6.18 kernel they are not necessary.

Monitoring Systems

With the new hardware, the following monitoring systems are available:

Frontend RAID card - Can be accessed from any web browser.
http://uscms1.fltech-grid3.fit.edu:888

NAS RAID card - Must be accessed from the frontend web browser.
http://nas-0-0:888

APC Primary UPS - Can be accessed from any web browser.
http://uscms1.fltech-grid3.fit.edu:3052