Instalación y configuración de un cluster con Rocks en la Universidad de Guadalajara

Autor	*Version*	*Fecha*
Adan Guerrero	0.1	16 de Enero de 2009

Tabla de Contenido

Índice de figuras

Introducción a un cluster de alto rendimiento

¿Qué es un cluster?

Cluster es un sistema de procesamiento de tipo paralelo o distribuido, que está formado de computadoras independientes, interconectadas entre sí, trabajando juntas como un solo recurso de cómputo intensivo.

Figura 1: ¿Porque un cluster?

Las caracteristicas más sobresalientes en la utilización de un cluster para el computo de alto rendmiento tiene las siguientes:

Elementos de un cluster

Procesadores

Se pueden utilizar practicamente cualquier tipo de procesadores. La tecnología actual los procesadores de una maquina accesible nos da un rendimiento similar a los procesadores de una supercomputadora. En donde cada procesador posee una gran cantidad de cache, así como de altas velocidades y bajo costo.

Comunicaciones

Existen soluciones que necesitan pocos recursos economicos para interconectar los equipos que formaran parte del cluster. Se puede utilizar cualquier tipo de tecnología para la interconexión entre los equipos ya sea la utlización de redes Ethernet, Myrinet, Gigabit. Con el que se obtiene un gran ancho de banda disponible para la comunicación con bajas latencias.

Sistemas Operativos

Se puede utilizar cualquier sistema operativo para la creación de un cluster sin embarlo se recomienda el uso de linux ya que este posee una gran estabilidad aúnado a un buen rendmiento en cuanto a manejo de memoría, así como de I/O eficiente, asi como la posibilidad de hacer un ajuste muy refinado a los parametros de los dispositivos para un mejor rendimiento.

Software

Existe una gran cantidad de software que ya esta listo para funcionar en un cluster, desde la aparición de los procesadores con HiperThreading (HT), la programación y la proliferación de software se ha desarrollado exponencialmente, con lo que se tiene una mayor cantidad de posibilidades para las diferentes disciplinas cientificas.

Recursos Humanos

El elemento más importante para el funcionamiento de cualquier sistema es el elemento humano que capacitado en la administración y manejo necesario de recursos proveé de un ambiente más amigable para aquellos usuarios que pretendan utilizar el cluster.

El cluster es facilmente escalable a comparación de las supercomputadoras en donde la escalabilidad depende de una gran cantidad de recursos economicos. Con la facilidad de extender el cluster con equipo de bajo costo la escalabilidad no representa una gran limitante en el momento de agregar recursos necesarios para incrementar el poder de computo.

Existen además muchas herramientas en la actualidad para la administración y manejo del cluster, tanto en herramientas de monitoreo, así como de herramientas para la administración de trabjos y recursos.

El soporte en librerias para pogramación en paralelo estan altamente desarrolladas, lo cual permite que la programción de nuevas aplicaciones que puedan funcionar en multiprocesamiento sea más sencillo.

Instalación de Rocks

(originalmente llamado NPACI Rocks) es una distribución de Linux para clusters de computadores de alto rendimiento. Fue inciada por la NPACI y la SDCS in 2000, y fue financiada inicialmente en parte por una subvención de la NFS (2000-2007), pero actualmente está financiada por la siguiente subvención de la NSF. Rocks se basó inicialmente en la distribución Red Hat Linux, sin embargo las versiones más modernas de Rocks están basadas en CentOS, con un instalador anaconda modificado, que simplifica la instalación 'en masa' en muchas computadoras.

Requerimientos y Prerrequisitos

Los requerimientos necesarios para instalar rocks son:

- Un conjunto de máquinas de arquitectura similar (compute nodes), cada una con una interfaz de red, disco duro con capacidad para más de 7 GB y memoria RAM superior a 256 MB.
- Un data switch (o varios) con un número de puertos mayor al doble del número de máquinas disponibles (para darle escalabilidad).
- Una máquina con 2 interfaces de red, capacidad en disco duro igual o superior a 20 GB, y memoria RAM superior o igual a 512 MB (frontend).
- Cables de red en número y longitud suficiente.
- Mueble o Rack con espacio apropiado para los chasis de las máquinas y eventualmente para el frontend, con acceso apropiado a la parte de atrás de los equipos.
- Una habitación con ventilación o refrigeración adecuada para los niveles de disipación de calor de todos los equipos combinados.
- Una UPS para alimentar al menos una máquina (el frontend por ejemplo) por más de 10 minutos.

Para la instalación del sistema operativo se debe disponer de los siguientes medios (rolls) que puede ser descargados desde el sitio de rocks clusters (http://www.rocksclusters.org/):

- Kernel Roll
- Core Roll
- OS Roll, disk 1
- OS Roll disk 2
- Cualquier otro Roll que considere necesario (Condor, Bio, Viz, etc.)

Instalación y configuración del Front End

Antes de proceder con la instalación del frontend es necesario asegurarse que las conexiones de la red externa y la red interna del cluster se hagan a la interfaz de red correcta. Rocks asume que la interfaz identificada como 'eth1' por el kernel será aquella que esta conectada a la red externa y la 'eth0' a la red privada del cluster.

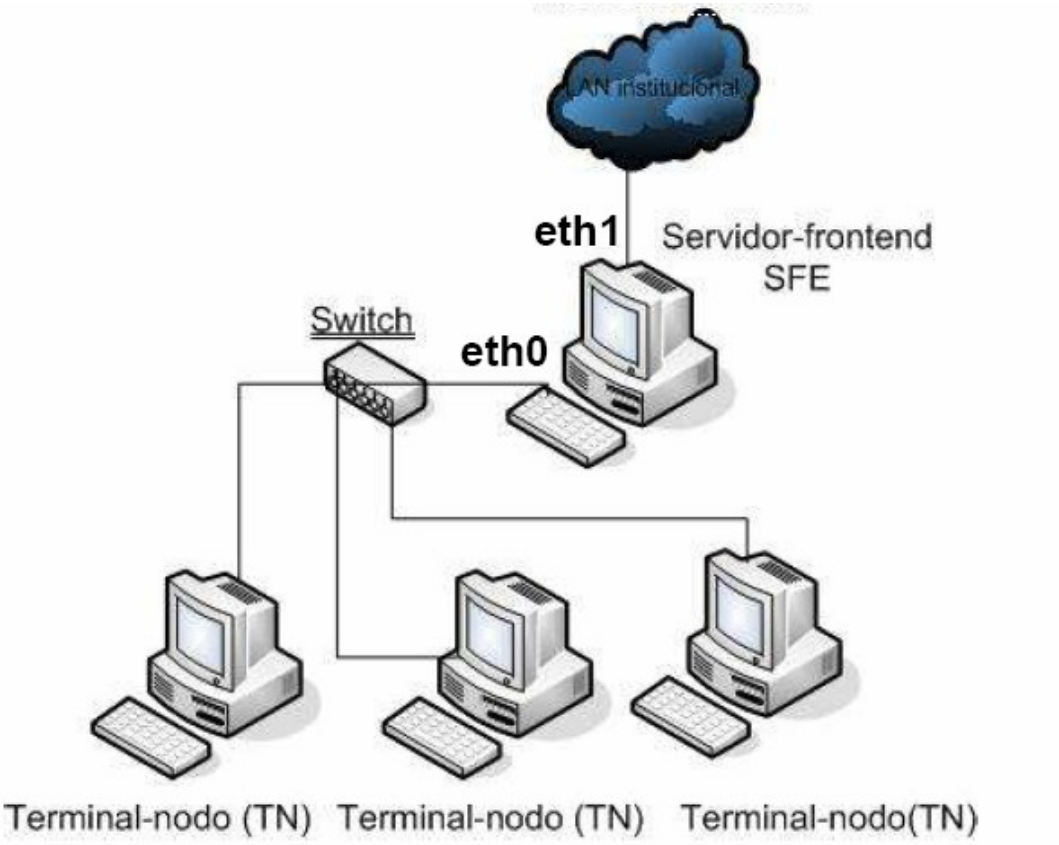


Figura 2: Esquema general de configuración fisica de un cluster

El proceso de instalación resulta muy sencillo una vez instalados los equipos en el rack y puesto la alimentación de poder hay que encender el equipo que será el nodo maestro (FrontEnd) y poner el disco de arranque (Boot) en la unidad de CD para comenzar la instalación. Al realizar este procedimiento aparecerá la primer pantalla en la que deberemos elegir la instalación del FrontEnd (Figura 3.).

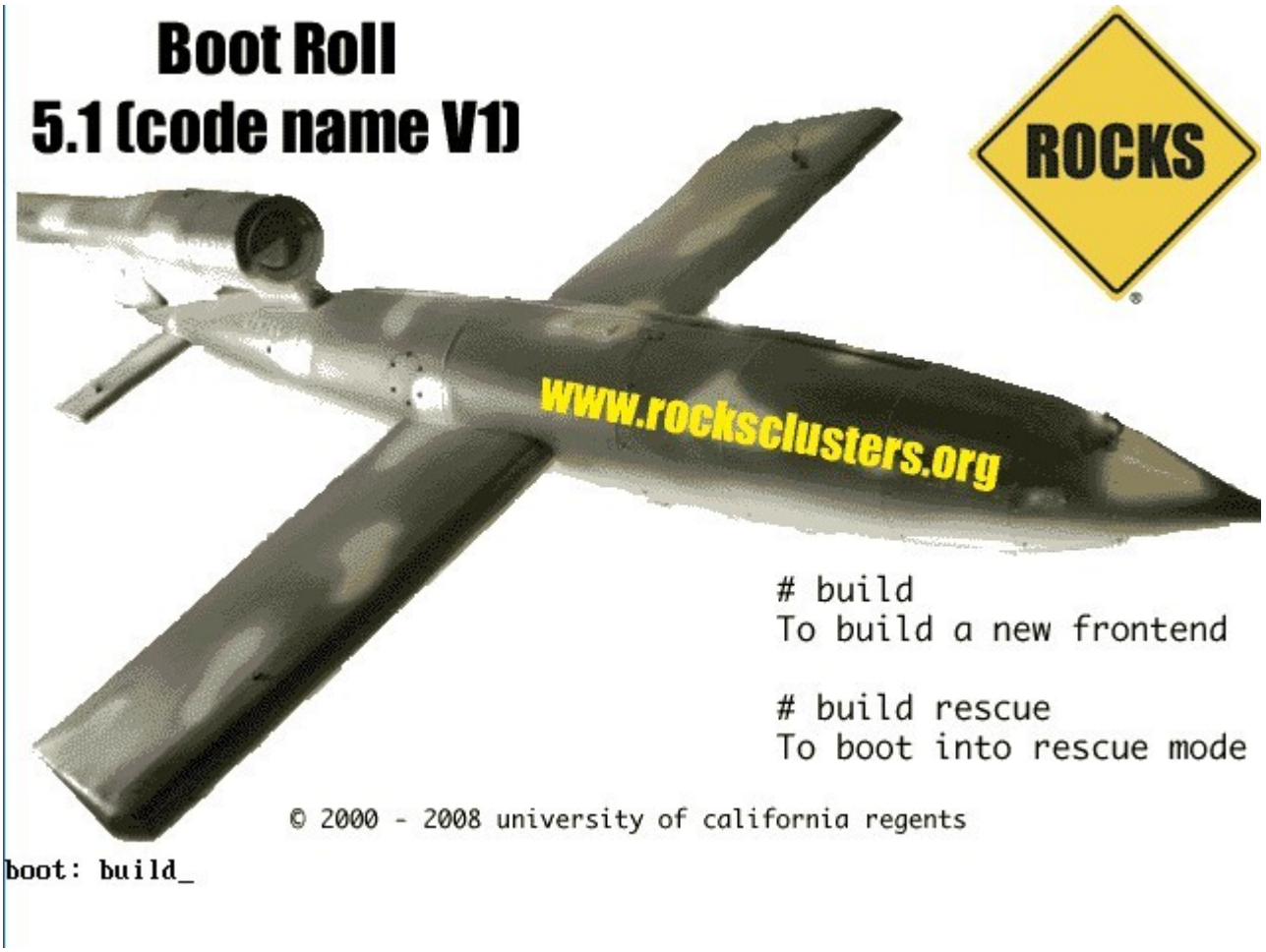


Figura 3: Pantalla de Inicio de la isntalación de Rocks

En esta pantala para poder instalar el nodo maestro debemos de escribir la palabra “build” (o frontend en versiones anteriores a la 5.1 de Rocks). Con esto comenzará el proceso de instalación en la que se configurará el cluster.

Pantalla 1. Seleccionar los rolls que se van a instalar: Para ello hay que selecionar la opcion ya sea de instalar desde CD o descargar desde un servidor. Para agregar los roll desde un CD solo hay seleccionar el roll a añadir de la pantalla que aparece despuede darle click en el boton “From media / From CD”, despues de selecionar el roll y presionar el boton “Submit” entonces regresamos a la pantalla anterior y la unidad de CD se expulsará automaticamente, despues de esto hay que ir repitienedo el mismo procedimiento para cada roll que se disponga. O en caso que se seleccione la opcion de instalar desde un servidor, seleccionar la lista de rolls que se desean instalar. Despues de esto presionar el Boton “Next”

Pantalla 2. Aparece una formulario en el que hay que llenar la información Básica del cluster entre las que se encuentran:

1. Fully Qualified Host Name (FQHN): hay que poner el nombre con el que se conocera en la red externa en este caso (orion.udg.mx)

2. Cluster Name: Nombre utilizado para identificar dentro de las herramientas del cluster como ganglia al cluster, se pueden utilizar cualquier tipo de combinación de letras y digitos (Orion).
3. Organización a la que pertenece el cluster, el estado y la ciudad (UdeG, Universidad de Guadalajara, Jalisco MX)
4. Dirección de contacto: Usar una dirección valida para referencia de correos electronicos respecto al cluster.

Pantalla 3. Configuración Básica de la red privada en nuestro caso una clase B (10.1.0.0 / 255.255.0.0)

Pantalla 4. Configuración de la red externa. En esta pantalla hay que dar los datos de la red en la que se encuentr nuestro cluster, en nuestro caso 148.202.105.223 / 255.255.255.0 gw 148.202.105.254.

Pantalla 5 Establecimiento de la contraseña de root que será utilizada para los propositos de adminsitración.

Pantalla 6. Establecer el uso horario del servido y en caso de ser requerido un servidor NTP para sincronizaciones de tiempo.

Pantalla 7. Esquema de particionado del disco, en caso de clusters dedicados se recomienda el autoparticionamiento por defecto que es el siguiente:

Partición	Punto de montaje	Tamaño por defecto	Tipo
a1*	/	7.7 GB	Ext3
a2	/var	3.9 GB	Ext3
a3	-	1 GB	Swap
a4	/state/partition1	Grow	Ext3

Figura 4: Esquema de particionamiento del disco

en donde “a” es valido tanto para sda como hda.

Una vez seleccionado el esquema de particionado comienza entonces la instalación del FrontEnd, en donde se nos solicitará ir agregando los cd de los rolls seleccionados o la descarga del servidor de los mismos. Una vez terminada la instalación de todos los rolls el equipo es reiniciado y el ultimo cd insertado es expulsado; una vez reiniciado el FrontEnd arranca con nuestro sistema Rocks ya instalado.

Configuración del Front End

Una vez instalado completamente el frontend se puede proceder a ajustar algunos detalles de la configuración de la máquina y la preparación de la distribución que será instalada en los compute nodes.
Para usuarios con teclados latinoamericano o español es muy conveniente cargar el mapa de teclado correspondiente. Para hacerlo se usa el comando:

```
# loadkeys /lib/kbd/keymaps/i386/qwerty/i386/es.map.gz
```

Para garantizar que la configuración del teclado se mantenga aún después de reiniciado el frontend se puede agregar esta línea al archivo rc.local:

```
# echo loadkeys /lib/kbd/keymaps/i386/qwerty/i386/es.map.gz >> /etc/rc.local
```

Para continuar con la configuración del frontend, antes de la instalación de los nodos de computo se recomienda completar las siguientes tareas:

1. Creación de cuentas de usuario
2. Preparación del /export/apps
3. Instalación de software desde tarballs
4. Instalación de paquetes desde RPM's
5. Preparación y creación de la distribución
6. Configurar el sistema 411 para que sincronice además de los archivos por default (I.e. etc/ld.so.conf)

Aunque no es necesario hacer todas estas tareas antes de instalar nuestro nodos es recomendable hacerlo. Una vez completado todo esto es momento de instalar nuestros nodos.

Instalación y configuración de los nodos

Para iniciar la instalación de los nodos es necesario entrar al FrontEnd e iniciar el proceso insert-ethers, y seleccionar la opción compute, que detectará a los nodos que se vayan a instalar.

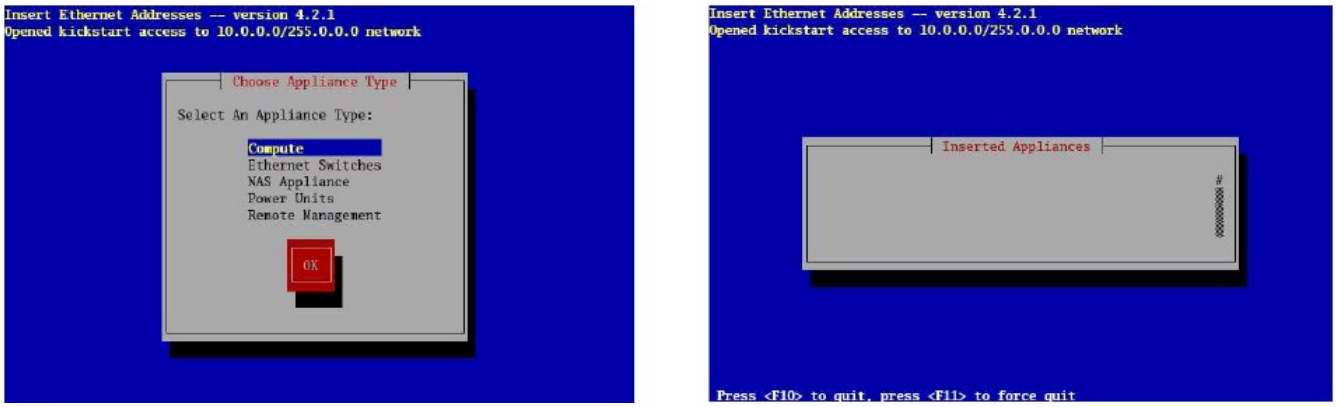


Figura 5: Comando insert-ethers

Despues de ejecutar el comando anterior hay que ir encendiendo los nodos con el cd de Kernel Roll para comenzar a instalar. Si se desea que se respete la ubicación física de los nodos y que coincida con la secuencia de asiganción de nombre hay que encender secuencialmente los nodos.

Se puede ir monitoreando el avance de la instalación de los nodos con el comando: rocks-console compute-0-0 reemplazando compute-0-0 por el nombre que le asigna insert-ethers aunque este comando solo funciona si se tiene una servidor X.

Configuración final de los nodos

Si la instalación de los nodos es exitosa no queda más que comenzar a utilizar el cluster. Pero se puede hacer una revisón previa para saber si todo funciona:

1. Verificar que el comando cluster-fork funciona correctamente
2. Revisara la salida de qstat -f
3. Un usuario puede conectarse exitosamente al frontend
4. Una vez dentro verificar si se puede conectar transparentemente a los nodos (no solictar contraseña)

Una vez realizadas estas pruebas podemos decir que nuestro cluster ya esta instalado.

Administración básica

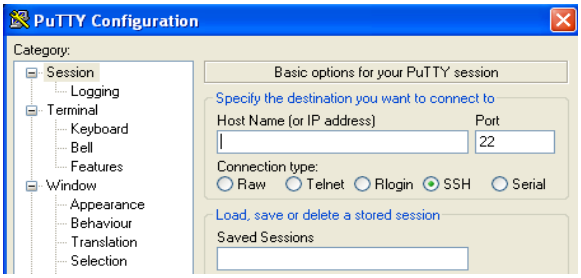
Acceso al Front-End y a los nodos del cluster

Acceso desde Linux

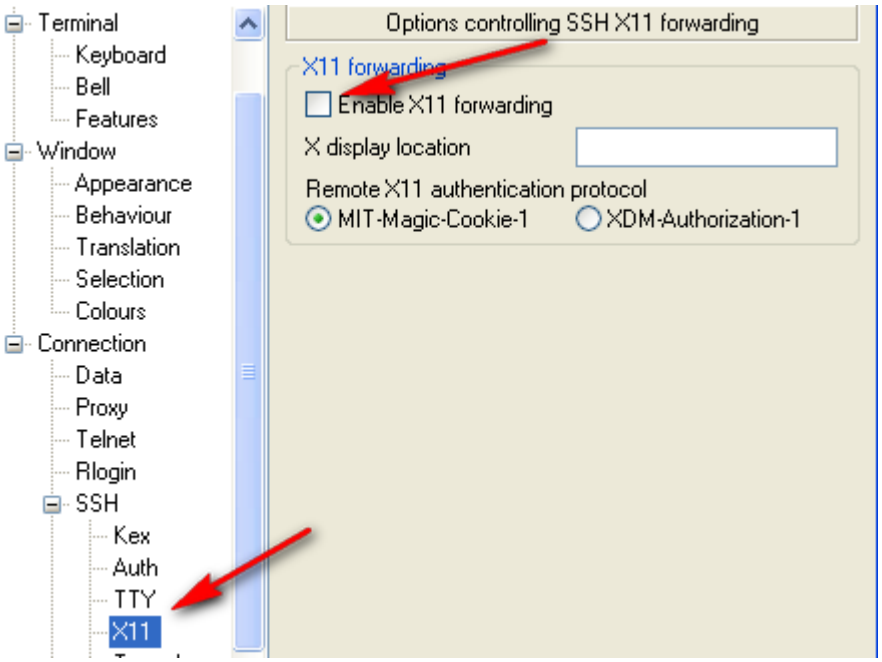
La mayoría de las máquinas Linux vienen dotadas de un cliente ssh que se invoca directamente desde el símbolo del sistema con el comando ssh.

```
$ ssh fulanito@cluster.dominio
$ ssh fulanito@192.168.3.2
Last login: Tue Oct 31 09:12:36 2006 from 192.168.3.2
Rocks 4.2.1 (Cydonia)
Profile built 00:06 13-Oct-2006
Kickstarted 19:36 12-Oct-2006
Rocks Frontend Node - Cluster
It doesn't appear that you have set up your ssh key.
This process will make the files:
/home/fulanito/.ssh/id_rsa.pub
/home/fulanito/.ssh/id_rsa
/home/fulanito/.ssh/authorized_keys
Generating public/private rsa key pair.
Enter file in which to save the key (/home/fulanito/.ssh/id_rsa):
Created directory '/home/fulanito/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/fulanito/.ssh/id_rsa.
Your public key has been saved in /home/fulanito/.ssh/id_rsa.pub.
The key fingerprint is:
30:8c:99:71:db:38:a2:91:99:9e:19:5d:ca:f7:01:c7 fulanito@cluster.dominio
```

Es también posible abrir una terminal remota usando un cliente ssh en Windows. Entre los clientes más populares, ligeros y versátiles para este sistema operativo se encuentran putty



La utilidad del putty es activar el X11 forwarding y tener instalado en nuestro equipo un servidor X para poder exportar aplicaciones en forma gráfica para ello solo hay que dar click en la opción de X 11 forwarding



Al activar el servidor X en nuestro equipo como el Cygwin (www.cygwin.com). Con esto podremos lanzar aplicaciones desde el fronend hacia nuestro equipo. Hay que tomar en cuenta que cualquier aplicacion gráfica que se utiliza hara uso de los recursos del frontend.

Sistemas de archivos en el cluster

Los archivos de un usuario son almacenados en su home directory (/home/fulanito) y estos debe de estar disponibles en cualquiera de los nodos del cluster cuando se establece una conexión con los mismos. Rocks utiliza dos mecanismos para que estos esten disponibles. Uno es el el sistema NFS (Network File Systems) con el que se comparte por medio de la red los recuros necesarios para que estos esten disponibles en cualquier equipo. El otro es el autofs que se asegura de montar el sistema de archivos NFS cuando un usuario se conecta a un equipo y se asegura de desmontarlo una vez que el usuario deja de utilizarlo.

Monitoreo básico de los recursos del cluster

Antes de comenzar a utiliza el poder de computo de un cluster se hace necesario conocer que recursos estan disponibles, basicamente existen tres recursos que son importantes para esto: a) el CPU, b) RAM y c)HD.

Una de las ventajas que tiene rocks es proporcionar una serie de herramientas para poder administrar el cluster, entre los que se encuentran la familia de comandos de rocks, en el que se encuentra el comando "cluster-fork" que nos permite envia la petición a todos los nodos o a un conjunto de ellos para ejecutar un comando en especifico sin tener que accesar a cada uno de los nodos. La sintaxis de este comando es:

cluster-fork [-hvm] [-p password] [-u host] [-d database] [-q sql-expr]

[-n nodes] [--help] [--list-rcfiles] [--list-project-info] [--verbose] [--bg]

[--verbose] [--rcfile arg] [--host host] [--password password] [--db database]

[--user host] [--query sql-expr] [--nodes encoded node list]

[--pe-hostfile sge machinefile] command

por ejemplo podemos hacer un ls desde el front end para saber si este comando funciona y todos los nodos responden adecuadamente

cluster-fork ls

cluster-fork –nodes compute-0-1 compute-0-2 ls

De esta manera podemos comenza a conocer nuestro cluster. Para el CPU nos interesan dos tipos de parametros, los estaticos y los dinamicos. Entre los estaticos que debemos consultar esta el numero de procesadores en la mother board, la velodidad del reloj y la chache. Para consultar estos parametro podemos enviar el siguiente comando:

cluster-fork cat /proc/cpuinfo <---- este archivo contiene información sobre el CPU aparecera algo como lo siguiente:

compute-0-8:

processor : 0

vendor_id : GenuineIntel

cpu family : 15

model : 2

model name : Intel(R) Xeon(TM) CPU 2.80GHz

stepping : 9
cpu MHz : 2791.114
cache size : 512 KB

processor : 1
vendor_id : GenuineIntel
cpu family : 15
model : 2
model name : Intel(R) Xeon(TM) CPU 2.80GHz
stepping : 9
cpu MHz : 2791.114
cache size : 512 KB

De los parametros dinamicos nos interesa la carga que tiene un nodo, para ello podemos realizar la consulta mediante el comando uptime.

cluster-fork uptime <--- con lo que obtenemos

compute-0-6:
18:51:55 up 6 days, 8:24, 0 users, load average: 0.00, 0.08, 0.27
compute-0-7:
18:51:56 up 6 days, 8:24, 0 users, load average: 0.02, 0.02, 0.08
compute-0-8:
18:51:57 up 6 days, 8:24, 0 users, load average: 0.00, 0.00, 0.00

en donde se puede estar consultado la carga promedio de cada uno de los nodos.

RAM, de este recurso tambien nos interesan tanto los parametro dinamicos como los estaticos. De los estaticos podemos ver el tamaño total de la memoria y el tamaño swap. De los dinamicos nos interesan la memoria utilizada y la memoria libre.

Para consultar los parametros dinamicos podemos revisar el archivo /proc/meminfo como sigue:

cluster-fork cat /proc/meminfo
MemTotal: 2595788 kB
MemFree: 2057760 kB

SwapTotal: 1020116 kB

SwapFree: 1020116 kB

Para el disco duro (HD) se pueden consultar las particiones del disco y los puntos de montaje, así como los tamaños asignados, el espacio disponible y el espacio utilizado, con el siguiente comando se puede realizar esta consulta:

cluster-fork df -kh

compute-0-7:

Filesystem Size Used Avail Use% Mounted on

/dev/sda1 5.7G 2.2G 3.2G 41% /

/dev/sda3 27G 77M 26G 1% /state/partition1

tmpfs 1014M 0 1014M 0% /dev/shm

orion.local:/export/home/adang

47G 21G 24G 46% /home/adang

compute-0-8:

Filesystem Size Used Avail Use% Mounted on

/dev/sda1 16G 2.2G 13G 16% /

/dev/sda2 3.8G 113M 3.5G 4% /var

/dev/sda5 13G 162M 13G 2% /state/partition1

tmpfs 1.3G 0 1.3G 0% /dev/shm

orion.local:/export/home/adang

47G 21G 24G 46% /home/adang

Otra herramienta de gran valor para el monitoreo es el comando ps, con el que podemos hacer una consulta de los procesos que se estan ejecutando en determinado momento, así como algunas de sus propiedades más importantes, un ejemplo de este comando es:

cluster-fork ps -caux

compute-0-7:

USER PID %CPU %MEM VSZ RSS TTY STAT START TIME COMMAND

root 1 0.0 0.0 2064 620 ? Ss Jan16 0:00 init

root 2 0.0 0.0 0 0 ? S< Jan16 0:00 migration/0

```
root         3  0.0  0.0      0     0 ?        SN   Jan16   0:00 ksoftirqd/0
root         4  0.0  0.0      0     0 ?        S<   Jan16   0:00 watchdog/0
root         5  0.0  0.0      0     0 ?        S<   Jan16   0:00 migration/1
root         6  0.0  0.0      0     0 ?        SN   Jan16   0:00 ksoftirqd/1
root         7  0.0  0.0      0     0 ?        S<   Jan16   0:00 watchdog/1
root         8  0.0  0.0      0     0 ?        S<   Jan16   0:00 migration/2
compute-0-8:
USER       PID %CPU %MEM    VSZ   RSS TTY      STAT START   TIME COMMAND
root         1  0.0  0.0   2064   620 ?        Ss   Jan16   0:00 init
root         2  0.0  0.0      0     0 ?        S<   Jan16   0:00 migration/0
root         3  0.0  0.0      0     0 ?        SN   Jan16   0:00 ksoftirqd/0
root         4  0.0  0.0      0     0 ?        S<   Jan16   0:00 watchdog/0
```

Organización del sistema Operativo

Rocks es una distribución basada en Redhat Enterprise Linux. La estructura de la distribución es por tanto similar en muchos aspectos a las distribuciones de ese mismo sabor (Fedora, CentOS, Scientific Linux, etc.) Las características especiales de trabajo en el cluster hacen sin embargo que hayan particularidades en la estructura del sistema operativo (servicios, sistemas de archivos, etc.)

1)La partición raíz (montada en el directorio '/') que contiene los archivos del sistema operativo y el espacio de almacenamiento de archivos temporales, logfiles, archivos de configuración,etc.

2) una partición especial para almacenamiento masivo en el frontend y en cada nodo que normalmente se monta sobre el directorio '/state/partition1'; en el frontend esta partición contiene las cuentas de usuario y otros archivos importantes relacionados con la instalación del sistema operativo; en los nodos esta partición puede usarse libremente para almacenar localmente grandes volúmenes de información.

3) en las últimas versiones de Rocks (>4.2) se ha incluido en el esquema de particionado por defecto una partición que se monta sobre el directorio '/var' que normalmente contiene información “variable” generada por los distintos programas y servicios del sistema operativo, incluyendo los logfiles.

De los sistemas de archivos locales en el frontend vale la pena resaltar los siguientes directorios de gran relevancia para el cluster:

– /export/home (/state/partition1/home):

```
total 68
drwxr-xr-x 3 condor condor 4096 Nov 1 14:22 condor
```

drwxr-xr-x 7 root root 4096 Nov 1 14:24 install
drwx------ 7 fulano fulano 4096 Nov 8 11:34 fulano

Este directorio contiene de un lado todos los home directory de los usuarios. De otra parte se encuentra allí también el directorio del usuario condor donde se depositan importantes archivos del sistema de colas de Condor.

– /export/home/install:
total 20
drwxr-xr-x 3 root root 4096 Nov 1 14:24 contrib
drwxr-xr-x 4 root root 4096 Nov 1 14:27 rocks-dist
drwxr-xr-x 13 root root 4096 Nov 1 18:58 rolls
drwxr-xr-x 3 root root 4096 Nov 1 15:37 sbin
drwxr-xr-x 3 root root 4096 Nov 1 19:12 site-profiles

Este importante directorio contiene la totalidad de los paquetes de instalación, archivos de configuración, programas y scripts especiales que usa Rocks para realizar la instalación del sistema operativo en los nodos.

Sistemas de archivos por red

Para montar automáticamente los directorios contenidos en /export sobre los nodos se configura el sistema autofs a través de los archivos /etc/auto.master, /etc/auto.home, / etc/auto.share. Normalmente estos archivos deben residir en el directorio /etc de todas las máquinas del cluster incluyendo el propio frontend.

Cuando se realizan cambios en los archivos de configuración del servicio autofs el servicio debe "recargarse". Esto se realiza usando el comando 'service' de linux como se ilustra a continuación:

service autofs reload
Checking for changes to /etc/auto.master ....
Reload map /usr/sbin/automount --timeout=1200 /share file/etc/auto.share
Reload map /usr/sbin/automount --timeout=1200 /home file /etc/auto.home

Se puede recargar el servicio también en otros (o todos) los nodos del cluster usando cluster-fork:

cluster-fork service autofs reload
compute-0-0:
Checking for changes to /etc/auto.master ....
Reload map /usr/sbin/automount --timeout=1200 /share file /etc/auto.share
Reload map /usr/sbin/automount --timeout=1200 /home file /etc/auto.home
compute-0-1:
Checking for changes to /etc/auto.master ....
Reload map /usr/sbin/automount --timeout=1200 /share file /etc/auto.share
Reload map /usr/sbin/automount --timeout=1200 /home file /etc/auto.home

El servicio 411 Secure Information System

Este servicio permite que archivos de configuración vitales para los servicios del cluster (listas de usuarios, tabla de passwords, grupos, configuración del servicio autofs, entre otras) sean compartidos por todas las máquinas de la plataforma, garantizando además (y ofreciendo las herramienta necesarias para) que se mantengan sincronizados a lo largo de la operación del cluster.

La lista de los archivos compartidos usando 411 puede encontrarse en el archivo /var/411/Files.mk

```
AUTOMOUNT = $(wildcard /etc/auto.*)

# These files all take a "#" comment character.
# If you alter this list, you must do a 'make clean; make'.
FILES = $(AUTOMOUNT)

# These files do not take a comment header.
FILES_NOCOMMENT = /etc/passwd \
        /etc/group \
        /etc/shadow

# FILES += /my/file
```

El servicio 411 esta configurado para realizar en forma automática la sincronización de los archivos de configuración en el cluster. Sin embargo en algunas situaciones es necesario "forzar" la sincronización después de que se ha hecho un cambio en los archivos de configuración (creación de un usuario, modificación de la configuración del servicio autofs, creación de un nuevo grupo, etc). La sincronización se puede realizar de tres maneras diferentes:

1. Usando service:
   Service 411 commit Este comando solo sincroniza los archivos que han cambiado
2. Usando make:
   make -C /var/411
   make -C /var/411 force ← Esto es similar solo que envía todos los archivos aunque no hayan sido modificados
3. usando 411get en todos los nodos
   cluster-for 411get ← Este mecanismo es util para detectar problemas de comunicación con los nodos.

Administración de Usuarios

Para agregar un usuario hay que relizar los siguientes pasos:

Creación de una cuenta de usuario

1. Creación básica de la cuenta: #useradd usuario
2. Asignación correcta del home del usuario: #usermod –d /home/usuario usuario
3. Asignación de la contraseña: passwd usuario
4. Configuración de autofs: vi /etc/auto.home agregar usuario cluster.local:/export/home/usuario
5. Sincronización de los archivos de usuario: $make –C /var/411 force
6. Recargar autofs en el frontend y los nodos: # service autofs reload ; cluster-fork autofs reload

En versiones más recientes lo anterior se puede reducir a tres pasos utilizando el comando rocks-user-sync

1. Creación de la cuenta
2. Fijación de la contraseña
3. Ejecución del comando: rocks-user-sync

Eliminación de cuentas de usuario

1. Eliminación de la cuenta. Para ello se usa el comando userdel: # userdel usuario
2. Desmontado del home directory: #umount /home/usuario; cluster-fork umount /home/usuario
3. Eliminación del home directory: #rm –rf /export/home/usuario
4. Sincronización de los archivos: Eliminar la entrada en /etc/auto.home

Síntesis de comandos

Comando	Explicación
# ls -ld /export	Muestra las propiedades del enlace simbólico / export
# exportfs -v	Muestra la lista de los sistemas de archivos compartidos vía NFS
# ls -l /export/	Revisa el contenido del directorio /export
# df -ht nfs	Muestra los sistemas de archivos tipo nfs montados en la máquina
# service autofs reload	Recarga el servicio autofs
# cluster-fork service autofs reload	Recarga el servicio autofs en todo el cluster
# service httpd status	Muestra el estado del servicio httpd
# service httpd start	Inicia el servicio httpd
# service 411 commit	Sincroniza los archivos de configuración que han cambiado recientemente
# make -C /var/411	Igual que el anterior
# make -C /var/411 force	Sincroniza TODOS los archivos de configuración
# cluster-fork 411get	Obtiene los archivos de configuración del frontend (operación en todos los nodos)
# useradd usuario	Crea una cuenta de usuario
# ls -al /export/home/usuario	Muestra el contenido (incluso el oculto) del directorio casa de usuario
# usermod -d /home/usuario usuario	Cambia el directorio casa del usuario en los archivos de configuración
# passwd usuario	Fija la contraseña de usuario
# ssh usuario@localhost	El root se conecta como usuario al frontend
# rocks-user-sync	Finaliza la creación de cuentas de usuario (se ejecuta después de useradd y passwd)
# usermod -g <gid> usuario	Cambia el número de grupo de un usuario
# userdel usuario	Borra la cuenta de usuario
# umount /home/usuario	Desmonta el home directory de usuario
# cluster-fork umount /home/usuario	Desmonta el home directory de usuario en todo el cluster
# rm -rf /export/home/usuario	Borra el home directory en el frontend

Tópicos Especiales de Administración

Instalación de Nuevo Software

La instalación de nuevo software en el frontend cluster se realiza siguiendo en principio procedimientos similares a los que se requieren para instalar software en cualquier servidor Linux. Sin embargo a la hora de requerir que el software pueda accederse desde todos los nodos, para ejecutarlo por ejemplo usando un Scheduler o para que las instancias de un programa en paralelo encuentren las componentes fundamentales del programa (bibliotecas, archivos de configuración, repositorios de temporales, etc.) es necesario configurar el paquete y los sistemas de archivos de manera apropiada.

1)instalación de un rpm de binarios

- # rpm -q gnuplot
- #cluster-fork rpm -q gnuplot
- # mkdir /export/apps/src
- # cluster-fork ls -l /share/apps
- cp -rf gnuplot-3.7.3-2.i386.rpm /export/apps/src
- # cluster-fork rpm -Uvh /share/apps/src/gnuplot-3.7.3-2.i386.rpm

Mas apropiado que instalar el rpm en caliente el paquete, es incluir el paquete directamente en la distribución que se instala en cada uno de los nodos. La ventaja evidente de este procedimiento estriba en el hecho que después de una re instalación de los nodos estará garantizado que el paquete se instale automáticamente sin requerir que se ejecuten las tareas descritas anteriormente. Para esto hay que seguir el siguiente procedimiento:

Rocks tiene un espacio especialmente dedicado a las contibuciones adicionales de los usuarios a la distribución instalada. El espacio esta habilitado en el directorio /export/rocks/install/contrib/<ver>/<arch>/

[root@orion i386]# ls
RPMS SRPMS
[root@orion i386]#

Allí se pueden colocar todos los archivos rpm que deseamos agregar a la distribución

por ejemplo:

cp <archivo>.rpm /export/rocks/install/contrib/5.1/i386/RPMS

Copiar ahi el rpm no es suficiente hay configurar la distribución para que incluya el paquete que acabamos de copiar, para ello hay que copar el archivo skeleton del directorio /export/rocks/install/site-profiles/<ver>/nodes a un archivos extended como sigue:

[root@orion nodes]# cd /export/rocks/install/site-profiles/5.1/nodes
[root@orion nodes]# cp skeleton.xml extended-compute.xml

Editar este ultimo y agregar los paquetes necesario a la seccion main, indicando SOLAMENTE el nombre del paquete (sin el número de la versión u otra información que venga con el archivo rpm) ie.

```
<!-- There may be as many packages as needed here. Just make sure you only
    uncomment as many package lines as you need. Any empty <package></package>
    tags are going to confuse rocks and kill the installation procedure
-->
<!-- <package>gnuplot</package> -->
<!-- <package> insert 2nd package name here and uncomment the line</package> -->
<!-- <package> insert 3rd package name here and uncomment the line</package> -->
```

Una vez configurado se debe reconstruir la distribución usando el comando rocks:

```
# cd /export/rocks/install
# rocks create distro
```

Es importante que una vez se reconstruye la distribución se pruebe al menos con la reinstalación de uno de los nodos que la distribución funciona correctamente. Para reinstalar fácilmente un nodo del cluster se puede recurrir a mecanismos de automatización que vienen instalados con Rocks y que usan el sistema kickstart. La re instalación procede de la siguiente manera:

a) Se elimina del nodo respectivo el archivo /.rocks-release

```
# ssh c0-6 rm -rf /.rocks-release
```

Esto habilita una opción en el gestor de arranque que hace que el nodo se reinicie la próxima vez en modo de re instalación.

b) Se inicia el proceso de re instalación:

```
# ssh c0-6 /boot/kickstart/cluster-kickstart
Shutting down kernel logger: [ OK ]
Shutting down system logger: [ OK ]
```

2) instalación de un tarball de fuentes

- # cp povray-3.6.tar.gz /share/apps/src
- # cd /export/apps/src
- # tar zxvf povray-3.6.tar.gz
- # cd povray-3.6.1
- # ./configure –prefix=/share/apps
- # make; make install
- Modificar /etc/profile en caso de ser necesario y recargar 411

3) instalación de una biblioteca de rutinas.

- # cp -rf gsl-1.6.tar.gz /export/apps/src

- # cd /export/apps/src
- # tar zxvf gsl-1.6.tar.gz
- # cd gsl-1.6
- # ./configure --prefix=/share/apps
- # make; make install
- Editar /etc/ld.so.conf:
- # ldconfig; cluster-fork ldconfig

Monitoreo de recursos por la web

Las tareas de monitoreo de los recursos del cluster que usamos en el documento 1 y muchas otras más pueden realizarse usando la interfaz web de Ganglia una poderosa y completa herramienta que viene instalada casi por defecto con todas las distribuciones de Rocks.

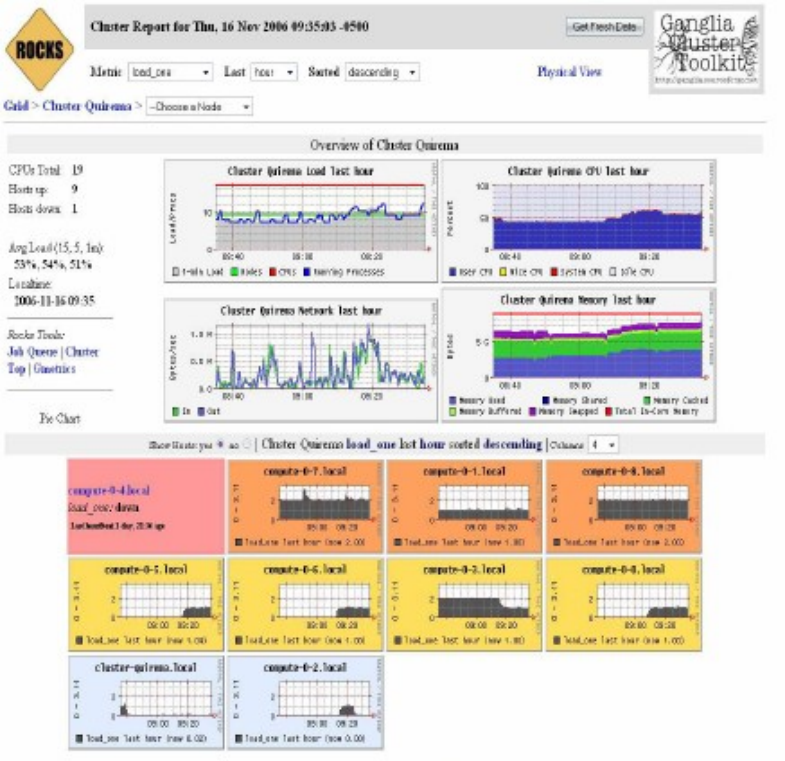


Figura 6: Ganglia monitoring

Solo hay que abrir un navegador y apuntar a la dirección IP o dns del cluster utilizando el protocolo https y agregando al final la palabra ganglia (https://orion.udg.mx/ganglia). Con lo que aparecera una pantalla como la figura 6.

Síntesis de comandos

Comando	Explicación
# ls -ld /export	Muestra las propiedades del enlace simbólico / export
# exportfs -v	Muestra la lista de los sistemas de archivos compartidos vía NFS
# ls -l /export/	Revisa el contenido del directorio /export
# df -ht nfs	Muestra los sistemas de archivos tipo nfs montados en la máquina
# service autofs reload	Recarga el servicio autofs
# cluster-fork service autofs reload	Recarga el servicio autofs en todo el cluster
# service httpd status	Muestra el estado del servicio httpd
# service httpd start	Inicia el servicio httpd
# service 411 commit	Sincroniza los archivos de configuración que han cambiado recientemente
# make -C /var/411	Igual que el anterior
# make -C /var/411 force	Sincroniza TODOS los archivos de configuración
# cluster-fork 411get	Obtiene los archivos de configuración del frontend (operación en todos los nodos)
# useradd usuario	Crea una cuenta de usuario
# ls -al /export/home/usuario	Muestra el contenido (incluso el oculto) del directorio casa de usuario
# usermod -d /home/usuario usuario	Cambia el directorio casa del usuario en los archivos de configuración
# passwd usuario	Fija la contraseña de usuario
# ssh usuario@localhost	El root se conecta como usuario al frontend
# rocks-user-sync	Finaliza la creación de cuentas de usuario (se ejecuta después de useradd y passwd)
# usermod -g <gid> usuario	Cambia el número de grupo de un usuario
# userdel usuario	Borra la cuenta de usuario
# umount /home/usuario	Desmonta el home directory de usuario
# cluster-fork umount /home/usuario	Desmonta el home directory de usuario en todo el cluster
# rm -rf /export/home/usuario	Borra el home directory en el frontend

Errores encontrados al instalar el cluster en la Universidad de Guadalajara

En esta sección se describirán los problemas encontrados al instalar el cluster.

Instalación de los nodos

Al instalar los nodos despues de tener el comando insert-ethers en el front end se encontraron dos errores:

1. Uno de los nodos no reconocía el disco, por lo que se tuvo que particionar manualmente desde un live cd para que el disco pudiera ser reconocido y particionado adecuadamente por rocks.
2. Al no terminar una instlación de manera exitosa, el front end no pudo borrar adecuadamente la entrada del dhcp, lo cual causaba un conflicto y no permitia la instalación del nuevo nodo, incluso utilizando la opción replace del insert-ethers, por lo que se tuvo que borrar manualmente la configuración del archivo /etc/dhcpd.conf y reiniciar el servicio de dhcp con el comando service dhcpd restart.

Reiniciar un servicio con errores

Algunos de los servicio puede caer en error, podemos comprobar este estatus con el comando

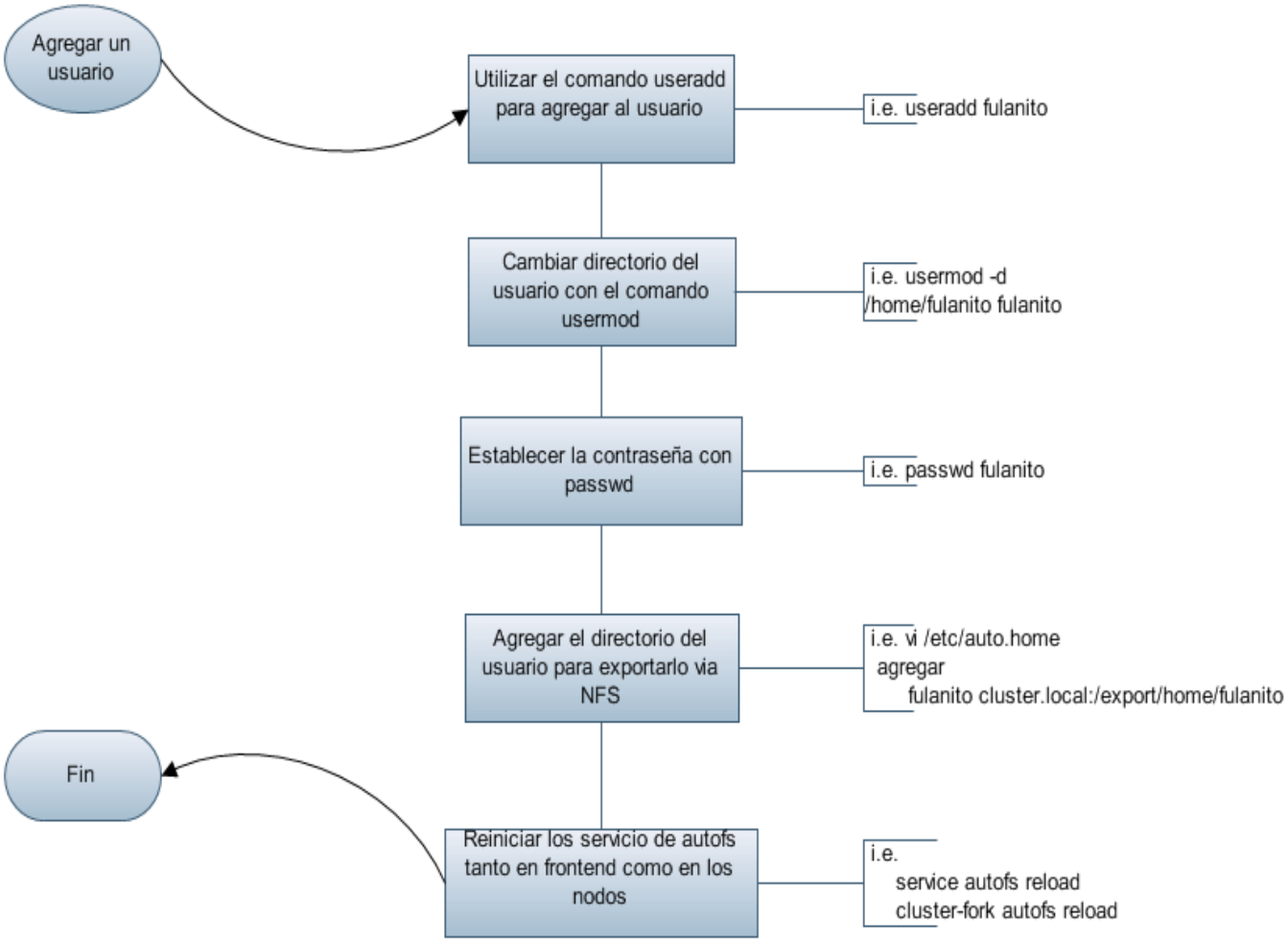
service <servicio> status i.e. service dhcpd status

Despues de revisar el estatus y verificar que este esta con errores hay que verificar porque tiene el error corregir y volver a inicar el servicio con

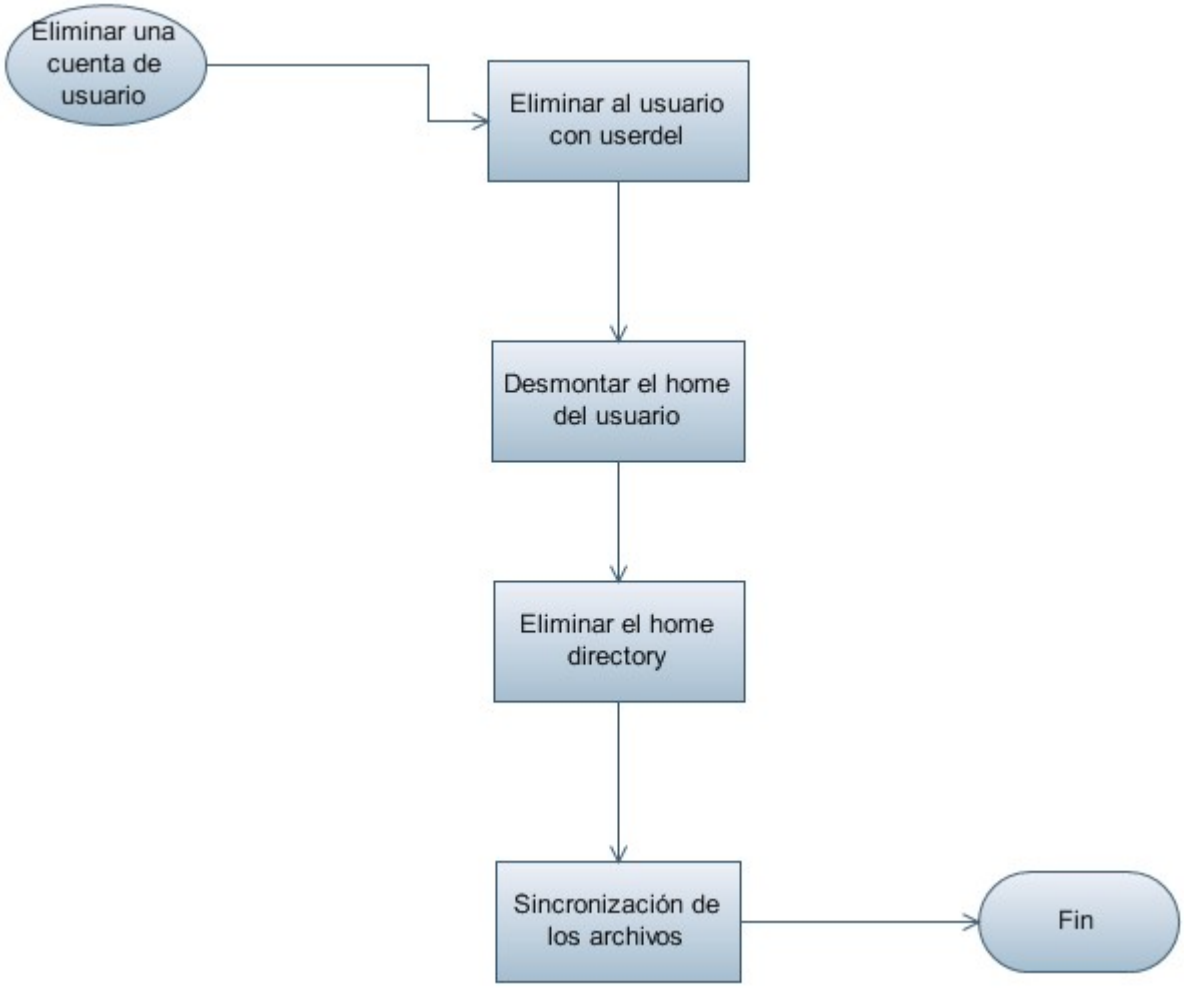
service <servicio> start i.e. service dhcpd start

Flujos para revisión de errores y checklist para la administración

Agregar usuario



Eliminar un Usuario



Detalles del cluster instalado en la Universidad de Guadalajara

La versión de rocks instalada en la Universidad de Guadalajara fue la 5.1 estable, con los rolls de:

- Kernel, Boot
- OS Disk 1
- OS Disk 2
- Core (area51+base+ganglia+hpc+java+sge+web-server+xen)

Se utilizaron los siguientes recursos de Hardware:

Switch:

Enterasys de 24 puertos dividido en 3 Vlans.

- Vlan ID 1 (default): Utilizada para dar salida a las maqinas, los puestos en esta vlan son del 1 al 5.
- Vlan ID 2 (Titan): En esta estan todos los puertos pares que es la utilizada para el cluster orion
- Vlan ID 3 (Atlante): En esta estan todos los puertos impares del 7 al 13

FrontEnd:

Hardware	Software
CPUs: 4 x 2.73 Ghz	OS: (x86)
Memory (RAM): 1.85 GB	Booted: January 16, 2009, 10:23 am
Local Disk: Using 32.303 of 70.090 GB	Uptime: 6 days, 9:56:04
Most Full Disk Partition: 48.5% used.	Swap: Using 60.8 of 996.2 MB swap.

Nodos:

compute-0-0

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 6.056 of 71.122 GB	Uptime: 6 days, 9:53:11
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-1

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 2.664 of 6.095 GB	Uptime: 6 days, 9:53:58
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-2

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 6.056 of 71.122 GB	Uptime: 6 days, 9:54:54
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-3

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 6.055 of 71.122 GB	Uptime: 6 days, 9:55:47
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-4

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 6.055 of 71.122 GB	Uptime: 6 days, 9:57:11
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-5

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 0.99 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 4.197 of 34.714 GB	Uptime: 6 days, 9:57:52
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-6

Hardware	Software
CPUs: 4 x 2.73 Ghz	OS: (x86)
Memory (RAM): 1.98 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 6.055 of 71.122 GB	Uptime: 6 days, 9:58:24
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute.0-7

Hardware	Software
CPUs: 4 x 2.73 Ghz	OS: (x86)
Memory (RAM): 1.98 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 4.197 of 34.714 GB	Uptime: 6 days, 9:59:00
Most Full Disk Partition: 43.7% used.	Swap: Using -0.0 of 996.2 MB swap.

compute-0-8

Hardware	Software
CPUs: 2 x 2.73 Ghz	OS: (x86)
Memory (RAM): 2.48 GB	Booted: January 16, 2009, 10:27 am
Local Disk: Using 4.395 of 34.260 GB	Uptime: 6 days, 9:59:23
Most Full Disk Partition: 19.6% used.	Swap: Using -0.0 of 996.2 MB swap.