Inquiry Cluster Administration

Christopher Dwan, Bioteam
First Delivered at IICB, Kolkata, India
December 14, 2009
Inquiry Cluster Administration

Apple system management tools
  – Workgroup manager: User accounts, groups, quotas
  – Server Admin: Software services
  – Software updates

Other software on the system
  – DSH, Sun Grid Engine, Ganglia

Cluster operating procedures
  – Power on / off, restart / monitor services

Backups and data recovery

Goal: Demonstrate the basic tools and best practices to enable local administrators to manage the system
Online Resources

Apple Server Documentation
  – http://apple.com/server/documentation

Sun Grid Engine
  – http://gridengine.info
  – http://gridengine.sunsorce.net
  – Mailing list: users@gridengine.sunsorce.net

Ganglia
  – http://ganglia.sourceforge.net/

Bioteam
  – http://faq.bioteam.net
  – support@bioteam.net
Contact information

Primary email support address  support@bioteam.net
Received by every member of the company
Tracked in a ticketing system

Websites
–  http://bioteam.net
–  http://faq.bioteam.net
  • Answers to common questions about inquiry and clustering
–  http://blog.bioteam.net
  • More detailed updates and posts on specific technologies of interest

My direct email:  cdwan@bioteam.net
Caveat:  I may be travelling and slower to respond than the ‘support’ alias.
Generic Portal Cluster Architecture

- Local Area Network
- Portal node(s)
- Private Network
- Compute Nodes

cdwan@bioteam.net
**Compute Cluster Management Principles**

- Nodes must be interchangeable
- Jobs should be scheduled through SGE whenever possible
- Three ‘R’s:
  - Reboot: Many problems can be solved by rebooting the offending node.
  - Re-image: Re-install the operating system to a clean state
  - Replace: Nodes that fail to work with an operating system reinstall have a hardware problem and should be replaced.

I no longer have any sense of humor for dealing with compute nodes on a one by one basis.

Might be tolerable with only 16 nodes, but more leads to madness
The OS X command line

Command line versions of GUI tools
  – Essential for making changes by remote, particularly on the compute nodes

• serveradmin
  – Start and stop www, nfs, etc.

• networksetup
  – Configure search domains, ip addresses, etc.

• systemsetup
  – Hostname, configuration stuff.

• http://apple.com/server/documentation
Server Admin

- GUI to manage system services
  - DHCP
  - DNS
  - NAT
  - LDAP
  - Firewall…
Workgroup Manager

• GUI to manage user accounts and groups
Server Monitor

- GUI to monitor detailed status of nodes
  - Temperature, fan speed, memory errors
- Connects through out of band “Lights out Management” interface
- Can boot and shutdown by remote.
DSH: Run command on all nodes

DSH: “Distributed shell”
http://www.netfort.gr.jp/~dancer/software/dsh.html.en

Run the same command on a set of hosts

Host list specified in /common/dsh/allhosts

dsh -a your_command argument1 argument2 ...

Verify time since reboot on all machines:

genesis2:named root# dsh -a uptime
executing 'uptime'
node001.cluster.private: 16:51 up 2:28
node002.cluster.private: 17:03 up 204 days
node003.cluster.private: 16:59 up 204 days
Procedures with dsh

Copy a file to /tmp on all machines:
Place the file in /common/scratch (or any shared location)
Copy from there to /tmp (or any local location):

```bash
dsh -a cp /common/scratch/the_file /tmp/the_file
```

Run software update on all machines:
```
dsh -a softwareupdate -i -a
```

Reboot all machines:
```
dsh -a shutdown -r now
```

Kill all mpiboot processes:
```
dsh -a killall -9 mpiboot
```
More DSH Examples

Check status of all SGE processes

dsh -a ps -elf \| grep sge
Password free ssh

For root, we configure password free ssh in Inquiry setup

- Node: /var/root/.ssh/authorized_keys
- Portal: /var/root/.ssh/id_dsa.pub

For users:

riptide:~ cdwan$ ssh-keygen -d
Generating public/private dsa key pair.
Enter file in which to save the key (/Users/cdwan/.ssh/id_dsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /Users/cdwan/.ssh/id_dsa.
Your public key has been saved in /Users/cdwan/.ssh/id_dsa.pub.

Many tools work more simply when password free ssh is configured.
Ganglia

Open source system monitoring tool
– http://ganglia.sourceforge.net/

gmond:
Every $N$ seconds, take system measurements (load, network I/O, etc)
Broadcast information
Listen for other broadcasts and record the data in /var/lib/ganglia/rrds

gmetad:
Provide web interface and draw graphs based on recorded data.

When ganglia reports nodes offline, usually the nodes are fine but ganglia needs a restart.
Ganglia Display
Restarting ganglia

Stop all ganglia processes on the portal
  SystemStarter stop GANGLIA

Stop ganglia processes on the nodes
  dsh -a SystemStarter stop GANGLIA

Start ganglia on the portal
  SystemStarter start GANGLIA

Start ganglia on the nodes
  dsh -a SystemStarter start GANGLIA
Sun Grid Engine

- Open source software for queuing and job control
- Currently distributed by Sun Microsystems.
- Resources:
  - http://gridengine.info
  - http://gridengine.sunsorce.net
  - users@gridengine.sunsorce.net
Sun Grid Engine Processes (summary)

Two processes on the portal:
- sge_qmaster
- sge_schedd

One additional process on the portal and also the nodes
- sge_execd

Each running job has an additional process:
- sge_shepherd
Verify that SGE is running

genusis2:named root# `qstat -f`

<table>
<thead>
<tr>
<th>queue name</th>
<th>qtype</th>
<th>used/tot.</th>
<th>load_avg</th>
<th>arch</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="mailto:all.q@genesis2.mit.edu">all.q@genesis2.mit.edu</a></td>
<td>BIP</td>
<td>0/2</td>
<td>0.39</td>
<td>darwin-ppc</td>
</tr>
<tr>
<td><a href="mailto:all.q@node001.cluster.private">all.q@node001.cluster.private</a></td>
<td>BIP</td>
<td>0/2</td>
<td>0.01</td>
<td>darwin-ppc</td>
</tr>
<tr>
<td><a href="mailto:all.q@node005.cluster.private">all.q@node005.cluster.private</a></td>
<td>BIP</td>
<td>0/2</td>
<td>-NA-</td>
<td>darwin-ppc</td>
</tr>
</tbody>
</table>

...
Restart SGE

On OS X, SGE processes are started and stopped via launchd. Kill them and they will instantly be re-started. Can make debugging difficult.

Stop all SGE processes on the portal and the nodes:

```
launchctl unload /Library/LaunchDaemons/net.sunsource.gridengine.sgeqmaster.plist
launchctl unload /Library/LaunchDaemons/net.sunsource.gridengine.sgeexecd.plist
dsh -a launchctl unload \
   /Library/LaunchDaemons/net.sunsource.gridengine.sgeexecd.plist
```

• Running jobs should survive this process

• SGE configuration is in /common/sge/default/common
• SGE logs are in /common/sge/default/spool
SGE Error States

‘au’: Alarm, Unreachable

- sge_qmaster process on portal cannot connect to the sge_execd process on the node
- Node could be offline, or the demon could just need to be restarted.

‘E’: Error

- Queues can be in error state
  qstat -explain E
- Jobs can be in error state
  qstat -j job_id
User Management: Add a user

- Connect using VNC, screen share, or remote desktop
- Open Workgroup Manager (Applications -> Server -> Workgroup Manager)
- Authenticate to LDAP domain (not local)
- Add user
  - Specify home directory in /Users

- Create user home directory from shell (does not happen automatically)
  
  ```
  sudo mkdir /Users/newuser
  sudo chown newuser /Users/newuser
  ```

- Create ssh keys for user
  
  ```
  sudo su newuser -
  ssh-keygen -d
  <return to accept defaults>
  Cp ~/.ssh/id_dsa.pub ~/.ssh/authorized_keys
  ```
Workgroup Manager: Add a user
Workgroup Manager: Home Directory
Workgroup Manager: Admin Authority

This user has full capabilities to make changes in Workgroup Manager.
Verifying a user account

On the portal:

```
genesis2:named root# id cdwan
uid=1027(cdwan) gid=20(staff) groups=20(staff),151
   (com.apple.sharepoint.group.4),80(admin),152
   (com.apple.access_ssh),150(com.apple.sharepoint.group.3)
```

On the nodes:

```
genesis2:named root# dsh -a id cdwan
executing 'id cdwan'
node001.cluster.private:  uid=1027(cdwan) gid=20
   (staff) groups=20(staff),80(admin)
node002.cluster.private:  uid=1027(cdwan) gid=20
   (staff) groups=20(staff),80(admin)
...
node013.cluster.private:  uid=1027(cdwan) gid=20
   (staff) groups=20(staff),80(admin)
```
Procedure: Power off cluster

1. Stop running jobs, log out users

2. Power down nodes
   ```
dsh -a shutdown -h now
   ```

3. Power down portal
   ```
   shutdown -h now
   ```

4. Power down attached storage

5. Power down network
Procedure: Power on cluster

• Power on network switch

• Power on external disk storage

• Power on portal
  – Log in and verify that the system is booted

• Power on compute nodes
  – Either using Server Monitor or the buttons.
  – Observe using ping, ganglia, and SGE
Operating System Services

Apple GUI
• Networking
• Domain Name Service (DNS)
• Dynamic Host Configuration Protocol (DHCP)
• Firewall
• Network Address Translation (NAT)
• Open Directory / LDAP

Command Line
• Automount
• NFS
Network configuration

• Portal:
  – Ethernet 1 = eth0: 192.168.2.254 / 255.255.255.0
  – Ethernet 2 = eth1: Public IP address

• Node00x
  – Ethernet 1: 192.168.2.x
  – Lights out Management: 192.168.2.10x admin / -secret-
  – Ethernet 2: Not connected
Domain Name Server (DNS)

Translates machine names to numeric IP addresses and back.

- All nodes and portal look to 192.168.2.254 for DNS
- DNS server on portal *forwards* to upstream DNS server

<table>
<thead>
<tr>
<th>Name</th>
<th>IP Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster.private</td>
<td>192.168.2.0 / 255.255.255.0</td>
</tr>
<tr>
<td>portal2net</td>
<td>192.168.2.254</td>
</tr>
<tr>
<td>node001</td>
<td>192.168.2.1</td>
</tr>
<tr>
<td>Node002</td>
<td>192.168.2.2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Node100</td>
<td>192.168.2.100</td>
</tr>
</tbody>
</table>
Network Health: DNS

- nslookup
  - google.com
  - <your portal name>
  - nslookup node001
  - nslookup node001.cluster.private
  - ping node001.cluster.private
  - ssh node001.cluster.private

- serveradmin status dns
DNS Configuration

DNS Configuration dialog box showing DNS settings for cluster.private network, including reverse and forward DNS records for various nodes. The dialog box also has options to add zones, records, and to stop and save DNS configurations.
DNS Configuration
Debugging DNS

Is the DNS service running?

genesis2:~ root# serveradmin status dns
dns:state = "RUNNING"

Is the DNS server entry set correctly?

genesis2:~ root# networksetup getdnsservers "Ethernet 1"
192.168.2.254

Can we resolve the private network by both full and partial names?
	nslookup portal2net.cluster.private
	nslookup portal2net
	networksetup getsearchdomains “Ethernet 1”

Can we resolve external hostnames?
	nslookup google.com
Debugging DNS

Restart the service:

```
genesis2:~ root# serveradmin stop dns
dns:state = "STOPPED"
genesis2:~ root# serveradmin start dns
dns:state = "RUNNING"
```

Log entries

```
Dec 11 13:54:21 genesis2 named[49343]: starting BIND 9.4.3-P1 -f
Dec 11 13:54:21 genesis2 named[49343]: command channel listening on 127.0.0.1#54
```

Configuration files in /var/named
Dynamic Host Configuration Protocol (DHCP)

Dynamic Host Configuration Protocol:
- At boot, client machines can broadcast a request for an IP address
- DHCP server replies with an offer of an IP address.
- The client either accepts or rejects that offer

Trouble signs with DHCP:
Network administrators and IT staff yelling at you
Cluster nodes boot, but do not appear on the network.

From /var/log/system.log

```
bootpd[49574]: DHCP REQUEST [en0]: 1,0:d:93:9d:39:8e <node001.cluster.private>
bootpd[49574]: domain search added
bootpd[49574]: replying to 192.168.2.1
bootpd[49574]: ACK sent node001 192.168.2.1 pktsize 385
```
DHCP Configuration
DHCP: Static Maps
Firewall – private side, permissive
Firewall – public side, restrictive
Network Address Translation (NAT)

All traffic from private network to public goes through NAT service on the portal

**Debugging:** If nodes cannot connect to outside servers, try restarting NAT:

```
serveradmin stop nat
serveradmin start nat
```

**Important note:**

NAT is implemented as a part of the firewall.
If the firewall is turned off, then NAT is not working.
Network Address Translation (NAT)
LDAP / Open Directory

- LDAP:
  - Lightweight Directory Access Protocol
  - Standard user authorization / authentication protocol

- We use a standard search path to make automatic installation simpler
  \[ dc=cluster,dc=private \]

User account management is separate from the service itself
  - Service is managed through Server Admin
  - User accounts are managed through Workgroup Manager
  - Must authenticate to the LDAP service within Workgroup Manager.
LDAP Configuration

Open Directory: Open Directory master

- LDAP Server: Running
- Password Server: Running
- Kerberos: Running

LDAP Search Base: dc=cluster,dc=private
Kerberos Realm: CLUSTER.PRIVARTE
Network File System (NFS)

Shared filesystems on the cluster, all exported via NFS

- /Volumes/data/common -> /common
- /Volumes/data/Users -> /Users
- /Library/Perl -> /RemotePerl

Mounts occur at boot time on the nodes via automounter.

On the nodes, /etc/auto_master includes the line:

/- /etc/bipod/auto.common

On the nodes, /etc/bipod/auto.common is:

/RemotePerl -ro,rsize=8192,wsize=8192 portal2net:/Library/Perl
/common -rw,rsize=8192,wsize=8192 portal2net:/Volumes/data/common
/Users -rw,rsize=8192,wsize=8192 portal2net:/Volumes/data/Users
Debugging NFS service on the portal

• Is NFS running?
  genesis2:named root# serveradmin status nfs
  nfs:state = "RUNNING"

• What volumes are exported?
  genesis2:named root# showmount -e
  Exports list on localhost:
  /Library/Perl                      192.168.2.0
  /Volumes/data                      192.168.2.0

• Can volume be mounted from the node?
  node001:~ root# mkdir test
  node001:~ root# mount 192.168.2.254:/common test
  node001:~ root# df -h test
  Filesystem Size   Used   Avail Capacity Mounted on
  192.168.2.254:/common 2.2Ti  931Gi  1.3Ti    42%    /private/var/root/test
Important directories / files
My debugging protocol
Backups / data protection / recovery

OTHER TOPICS
**Important directories / files**

- **OS log file:** /var/log/system.log
- **Inquiry log file:** /common/scratch/inquiry.log
- **SGE log file:** /common/sge/default/spool/qmaster/messages
- **Web root:** /Library/WebServer/Documents
- **PERL (cluster wide):** /RemotePerl
General Cluster Health

- **Network**
  - Portal and nodes resolve themselves and each other
  - Cluster.conf in dns search paths
  - Is the Firewall on? Should it be? Does turning it on/off resolve the problem?

- **Shared directories**
  - NFS shares exported from portal, mounted by nodes

- **User authentication** is working from portal to nodes

- **SGE** on portal, then on nodes
Backups and data recovery

• Snapshot backup:
  – Protects against disk failure
  – Bring system back online to a known-good configuration
  – Minimize downtime

• Incremental backups:
  – Traditional daily / weekly / monthly backups
  – User data, files that change.

• Archival data storage:
  – Long term storage
  – Data does not change, but must be maintained
  – Possibly offline
Snapshot backups for OS protection

- **Tier 1: Internal RAID**
  - Redundant Array of Independent Disks (RAID)
  - RAID 1: Mirror set (two or more disks kept identical by the operating system)
  - Single disk failure:
    - Replace failed disk, rebuild RAID
  - Double disk failure before RAID rebuild is complete:
    - System is offline, loss of all changes since Tier 2

- **Tier 2: Snapshot**
  - Create a bootable image of a known good configuration on a USB drive
  - Place that USB drive on a shelf.
  - Recovery:
    - Repair hardware (replace disks, new motherboard, …)
    - Copy image from USB drive to boot disk of server
Bad idea for long term archives
Incremental backups

Goals:
– Return to earlier version of file.
– Protect against unwanted changes, corruption, or deletion
– Do not waste time and disk redundantly storing the same data

Traditional daily / weekly / monthly setup
– Keep daily changes for a week
– Keep weekly changes for a month
– Keep monthly backups until disk is full

Options:
– Built in: Apple’s Time Machine
– Commercial: Retrospect  http://www.retrospect.com/
Archival data storage

• Backups are a nightmare, because components always fail
• Data loss is costly in time, reputation, and morale
• Most major groups use a tiered system:
  – RAID protection for all disks
  – Double parity (RAID 6) is essential for large data stores
  – Mirrored disks (RAID 1) are essential for key servers
  – First level backup is disk to disk
  – Archival data is written to tape, once, and driven offsite
Questions