Training for University of Richmond

advanced clustering
technologies, inc.

2

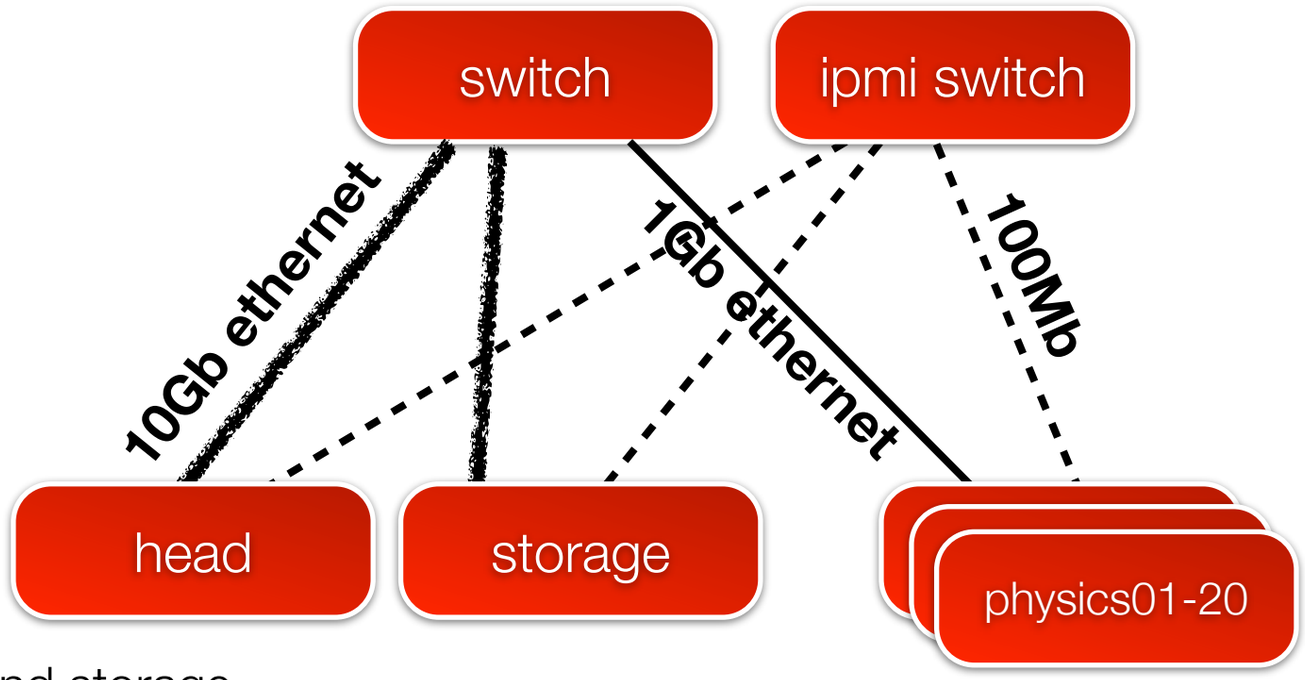
Agenda

- Cluster Overview
- Software
  - Modules
  - PBS/Torque
  - Ganglia
  - ACT Utils

advanced clustering technologies, inc.

3 Cluster overview

- Systems
  - 1x head node
  - 1x storage node
  - 20x compute nodes
- Network
  - 1Gb Ethernet to nodes
  - 10Gb Ethernet to head and storage
  - All nodes connected to dedicated IPMI management network



advanced clustering technologies, inc.

4

Network

- Private network for node to node communication
  - 10.1.1.0/24
- Private network for IPMI communications
  - 10.1.3.0/24
- Head node is only machine with “public” IP address, provides firewall to protect cluster network

advanced clustering technologies, inc.

5

Cluster overview

- Nodes
  - Dual CPU 6-core Intel Xeon “Westmere” processors at 2.66GHz
  - 24GB of RAM
  - Pair of 500GB drives in RAID0 striped configuration
- Head
  - Dual CPU 6-core Intel Xeon “Westmere” processors at 2.66GHz
  - 24GB of RAM
  - 4x 500GB drives in a RAID10
  - NFS exports /act filesystem to all nodes

advanced clustering technologies, inc.

6

Cluster overview

- Head continued
  - Runs Torque scheduler and server process
  - DHCP server for cluster network
  - Webserver for ganglia monitoring server
- Storage
  - Single CPU 6-core Intel Xeon “Westmere” processors at 2.66GHz
  - 12GB of RAM
  - 8x 1TB drives in a RAID5 + hotspare
  - NFS exports /home filesystem to all nodes

advanced clustering technologies, inc.

7

Modules command

- Modules is an easy way to setup the user environment for different pieces of software (path, variables, etc).
- Setup your .bashrc or .cshrc
  - source /act/etc/profile.d/actbin.[sh|csh]
  - source /act/Modules/3.2.6/init/[bash|csh]
  - module load null

advanced clustering technologies, inc.

8

Modules continued

- To see what modules you have available:
  - module avail
- To load the environment for a particular module:
  - module load modulename
- To unload the environment:
  - module unload modulename
  - module purge (removes all modules from environment)
- Modules are stored /act/Modules/3.2.6/modulefiles - can customize for your own software

9

PBS/Torque introduction

- Basic setup
- Submitting serial jobs
- Submitting parallel jobs
- Job status
- Interactive jobs
- Managing the queuing system

advanced clustering technologies, inc.

10

Basic setup

- 3 main pieces
  - pbs_server - main server components responds to user commands, etc
  - pbs_sched - decide where to run jobs
  - pbs_mom - a daemon that runs on every node that will execute jobs
- Each node has 12 slots which can be used for any number of jobs
- There is currently only 1 queue, named “batch” it’s set as a FIFO (first-in first-out)

advanced clustering technologies, inc.

11

Submitting batch jobs

- Basic syntax: qsub jobscript
- jobscripts are simple shell scripts in either SH or CSH which at a minimum contain the name of your program. Here is the minimum jobscript:

```
#!/bin/bash
/path/to/executable
```

12

Common qsub arguments

- -q queuename
  - name of the queue to run the job in
- -N jobname
  - a descriptive name of the job
- -o filename
  - path to the filename to write the contents of STDOUT
- -e filename
  - path of the filename to write the contents of STDERR

13

Common qsub arguments

- -j oe
  - Join the contents of STDERR and STDOUT into one file
- -m [a|b|e]
  - Send out e-mail at different states (a = job aborted, b = job begins, e = job ends)
- -M emailaddr
  - email address to send messages to
- -l resourcename=value,[resourcename=value]
  - a list of resources needed to run this job

14

Resource options

- walltime
  - maximum amount of real time the job can be running (if exceeded it will be terminated)
- mem
  - maximum amount of memory to be consumed by the job
- nodes
  - number of nodes requested to run this job

advanced clustering technologies, inc.

15

Submitting serial jobs

- A moderately complex job script can suggest command line parameters to PBS (prefixed with #PBS) that you may have left off of qsub as well as perform environment setup before running your program:

```
#!/bin/bash
#PBS -N testjob
#PBS -j oe
#PBS -q batch

echo Running on `hostname`.
echo It is now `date`.
sleep 60
echo It is now `date`.
```

16

Submitting parallel jobs

- Very similar to batch jobs except a new argument “-l nodes=X:ppn=X”
  - nodes: number of physical servers to run on
  - ppn: processors per node to run on, i.e. 12 to run on all 12 cores
- Examples:
  - run on 2 nodes using 12 cores per node, for a total of 24 cores: -l nodes=2:ppn=12
  - run on 4 nodes using 1 core per node, and 2 nodes using 2 cores per node: -l nodes=4:ppn=1+2:ppn=2

17

Job status

- You can check your own job submission status by looking at the output of "qstat". qstat only shows your own jobs, be default.
- To show jobs for all users, run "qstat -u '*'".
- To examine the details of a jobs, use "qstat -f jobid"
- Common job states
  - R = running
  - Q = queued
  - E = error

18

Interactive jobs

- Using qsub -I you can submit an interactive job.
- When a job is scheduled, it lands you in a shell on the remote machine
- You can pass any argument that you'd normally pass to qsub (i.e. qsub -N name -I nodes=1:ppn=5).
- When you exit, the resources are immediately freed for others to use.

19

Managing the queuing system

- qdel - delete a job that has been submitted
- qalter - alter a job after submission
- qhold - hold a job in the queue and do not execute
- qrls - release a hold on a job
- pbsnodes - see nodes configured in the system
- pbsnodes -o nodename - take a node offline from the queuing system
- pbsnodes -c nodename - clear the offline state of a node
- qmgr - create queues and manage system properties

20

More information

- Administrator manual:
  - http://www.clusterresources.com/products/torque/docs/

advanced clustering
technologies, inc.

21

Ganglia

- Ganglia installed on the master node and available at
  - http://test3.advancedclustering.com/ganglia/
- gstat command available - provides a command line overview of the ganglia collected data

22

ACT Utils

- ACT Utils is a series of commands to assist in managing your cluster, the suite contains the following commands:
  - act_authsync - sync user/password/group information across nodes
  - act_cp - copy files across nodes
  - act_exec - execute any Linux command across nodes
  - act_netboot - change network boot functionality for nodes
  - act_powerctl - power on, off, or reboot nodes via IPMI or PDU
  - act_sensors - retrieve temperatures, voltages, and fan speeds
  - act_console - connect to the hosts's serial console via IPMI

23

ACT Utils common arguments

- All utilities have a common set of command line arguments that can be used to specify which nodes to interact with
  - --all all nodes defined in the configuration file
  - --exclude a comma seperated list of nodes to exclude from the command
  - --nodes a comma separated list of node hostnames (i.e. physics01,physics02)
  - --groups a comma separated list of group names (i.e. nodes)
  - --range a “range” of nodes (i.e. physics01-physics05)
- Configuration (including groups and nodes) defined in /act/etc/act_utils.conf

24

Groups defined on your cluster

- nodes - all compute nodes

advanced clustering
technologies, inc.

25

ACT Utils examples

- Find the current load on all the compute nodes
  - act_exec -g nodes uptime
- Copy the /etc/resolv.conf file to all the login nodes
  - act_cp -g nodes /etc/resolv.conf /etc/resolv.conf
- Shutdown every compute node except physics01
  - act_exec --group=nodes --exclude=physics01 /sbin/poweroff
- tell nodes physics01 - physics03 to boot into cloner on next boot
  - act_netboot --nodes=a1pcmp01,a1pcmp03 --set=cloner-v3.14

26

ACT Utils examples

- Check the CPU temperatures on all nodes
  - act_sensors -g nodes temps
- Connect to the console of physics05
  - act_console --node=physics05
  - If connected with X11 forwarding:
    - act_console --use_xterm --node=physics05
- Hardware power control
  - act_powerctl --group=nodes [on|off|reboot|status]

27

Shutting the system down

- To shut the system down for maintenance:
  - act_utils -g nodes --node=storage /sbin/poweroff
- Then shut down the head
  - /sbin/poweroff