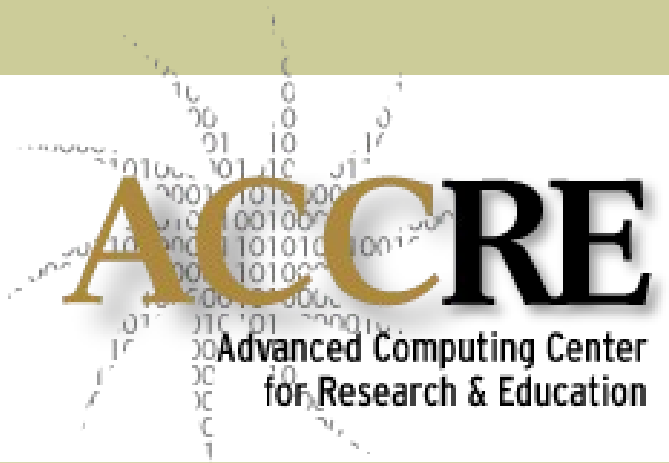

Introduction to the Cluster

http://www.accre.vanderbilt.edu

2008 April

ACCRE Advanced Computing Center for Research & Education

Agenda

- Resource overview (slides 3-6)
- Logging on with `ssh` and X tunneling (slide 7-8)
- Transferring files to and from the cluster with `scp` (slide 9)
- Setting your environment and `setpkgs` (slides 10-14)
- Scheduler basics and ACCRE policies (slides 15-16)
- Requesting resources - submitting jobs (slides 17-20)
- Checking on submitted jobs (slides 21-25)
- Cluster etiquette - running jobs (slides 26-28)
- Cluster disk space and etiquette (slides 29-30)
- ACCRE storage policies (slides 31)
- Getting help (slide 32)

The Cluster

ACCRE Advanced Computing Center for Research & Education

Cluster X86 Processors

- **~760 cores in dual or dual-dual nodes (faster floating point operations)**
- **~220 nodes /440 cores, 2.0GHz AMD (dual) Opterons**
  - 180 nodes with 1 GB memory
  - 60 nodes with 2 GB memory
  - 50% with Myrinet networking
- **80 nodes / 320 cores, 1.8GHz / 2.4 Ghz AMD (dual dual) Opterons**
  - 80 nodes with 4 GB memory
  - 100% Ethernet networking

Cluster PowerPC Blades

- 644 JS20 IBM PowerPCs in dual blades (faster integer operations)
- 322 blades with 2.2GHz PowerPC processors
- 1.5GB memory
- ~50% with Myrinet networking

ACCRE Advanced Computing Center for Research & Education

Cluster Details

- Each brood: 20 x86 + gateway or 28 PPC blades + gateway
- Communication between broods (groups of compute nodes or blades) and gateways and outside world is 1Gbps bandwidth Ethernet
- Connection between compute nodes are Ethernet or 2Gbps low-latency Myrinet (for parallel apps that can take advantage)
- For more details see the High Performance Compute Cluster page at our web site:

http://www.accre.vanderbilt.edu/mission/services/hpc.php

Logging On

- SSH (Secure Shell)
  - `ssh username@vmplogin.accre.vanderbilt.edu`
  - `ssh username@ppclogin.accre.vanderbilt.edu`
  - Round robin to one of the sires/gateways to distribute load on gateways
- Changing your password on ***vmpsched***
  - `ssh username@vmplogin.accre.vanderbilt.edu`
  - `ssh username@vmpsched`
  - `passwd`

X Windows Remote Display

- Displaying graphics over net can be slow
- Run X server
- Turn on **ssh** X11 tunneling when connect, e. g., for OS X and Linux:

```
ssh -X user@vmplogin.accre.vanderbilt.edu
```

- Set up directions, also for Windows, see:

  www.accre.vanderbilt.edu/support/selfhelp/faq.php#xremotedisplay

- Example

Transferring Files To/From Cluster

- GUI SSH client: http://www.ssh.com
- Command line Secure Copy – **scp**
  - Usage like Unix "**cp *file1 file2***" (source to destination)
  - But can use to transfer files between remote machines, e. g.,

If on cluster, to copy from outside machine (i.e. your desktop)

```
scp username@outsidemachine:file /your/cluster/dir
```

If on outside machine, to copy to cluster

```
scp -r /some/dir/* username@vmplogin:/your/cluster/dir
```

- Also **sftp**

Your Environment

- ***.bashrc*/*.bash_profile*** (for ***bash***)
  - **export** ***env_variable=definition***
  - **export** ***PATH=/home/username/bin:$PATH***
  - **setpkgs -[aer]** ***package_name***
- ***.cshrc*** (for ***csh*** or ***tcsh***)
  - **setenv** ***env_variable "definition"***
  - **setenv** ***PATH"/home/username/bin:$PATH"***
  - **setpkgs -[aer]** ***package_name***
- E. g., add **/usr/lpp/mmfs/bin** to $PATH

setpkgs / pkginfo

- Usage:
  - `setpkgs` with no options prints help to screen (no man page)
  - `setpkgs -a package_list` adds environment variables
  - `setpkgs -e package_list` erases environment variables
  - `setpkgs -r package_list` replaces all with packages listed
  - `pkginfo` with no options prints list of installed packages
  - `pkginfo -p package -i` prints detailed info on ***package***
- Examples

setpkgs / pkginfo

- Can auto-set cluster environment depending on machine architecture by adding to your login files:
  - ***.bashrc*** (or ***.bash_profile***):

```
if [`arch` == "ppc64" ]; then
   #Put your ppc64 statements here
   #E.g., setpkgs commands
   setpkgs -r
else
   #Put your x86_64 statements here
   #E.g., setpkgs commands
   setpkgs -r
fi
```

setpkgs / pkginfo

- *.cshrc*:

```
if (`arch` ==“ppc64” ); then
   #Put your ppc64 statements here
   #E.g., setpkgs commands
   setpkgs -r some_pkg
else
   #Put your x86_64 statements here
   #E.g., setpkgs commands
   setpkgs -r some_pkg
endif
```

setpkgs / pkginfo

- **Example:**

```
if [ `arch` == "ppc64" ]; then
        echo "in ppc"
        setpkgs -a openmpi_gcc-ibm_ether
        setpkgs -a gcc_compiler
        export ARCHPATH=$HOME/ppc64
        NODETYPE=powerpc
else
        echo "in x86_64"
        setpkgs -a openmpi_gcc_ether
        setpkgs -a gcc_compiler
        export ARCHPATH=$HOME/x86_64
        NODETYPE=intel
fi
export NODETYPE
```

Scheduler Basics

- Scheduling jobs (slide 16)
  - www.accre.vanderbilt.edu/mission/cluster_policies/job_scheduler.php
- `qsub` and PBS scripts (slides 17-19)
  - http://www.accre.vanderbilt.edu/support/selfhelp/gettingstarted.php
  - http://www.accre.vanderbilt.edu/support/selfhelp/faq.php
  - `man qsub ; man pbs_resources`
- Resources available:
  http://www.accre.vanderbilt.edu/mission/services/hpc.php#nodes
- Using the scheduler (slides 20-27)

How The Scheduler Works

- Submit jobs to the scheduler
  - `qsub [options] PBS_script`
- TORQUE/PBS resource manager - PBS MOM (machine oriented miniserver) runs on nodes executes instructions, keeps track of resources and usage
- Maui/Moab job scheduler - gets resources from PBS and schedules jobs based on:
  - Fairshare contribution – from CPU buy-in
  - Job run priority – calculated based on ~80% fairshare usage and ~20% queuetime

ACCRE Advanced Computing Center for Research & Education

PBS Script

```
#!/bin/tcsh
#PBS -M my.address@vanderbilt.edu
#PBS -m bae
#PBS -l nodes=4:ppn=2:x86
#PBS -l walltime=00:30:00

#PBS -l mem=1000mb

#PBS -o myjob.output
#PBS -j oe
```

- first line defines shell
- send status/progress emails
- email at beginning, abort, & end
- resources (`-l`) required for job
- REQUIRED! estimated wall clock (hh:mm:ss or ssss.ss)
- max=node mem minus ~200mb; lower limit=~1mb; default=400mb
- send stdout to *myjob.output*
- join stdout/err to *myjob.output*

```
echo "This is my first job submitted to the ACCRE cluster."
# Script comment: replace echo with your script/executables
# resource list can be complicated for parallel codes
# node attributes defined by our specific hardware (slide 19)
```

ACCRE Advanced Computing Center for Research & Education

PBS Script Example

```
#!/bin/sh
# Resource list
#PBS -l nodes=1:ppn=1:x86
#PBS -l walltime=15:00
#PBS -l cput=15:00
#PBS -j oe
# Defining environment variables for convenience
# Name of your Matlab script
PROGRAM="~/test/matlab/matlab.script"
# Save output to file output.txt
OUTPUT="~/output.txt"
# This is the equivalent of: /usr/local/matlab/bin/matlab <
   ~/test/matlab/matlab.script > ~/output.txt

matlab < $PROGRAM > $OUTPUT
```

ACCRE Advanced Computing Center for Research & Education

qsub Node Attributes

- Cluster specific qsub/PBS node attributes
  - *ppc64, nomyrinet*
  - *ppc64, myrinet*
  - *x86, opteron, nomyrinet*
  - *x86, opteron, nomyrinet, bigmem*
  - *x86, opteron, nomyrinet, dualdual*
  - *x86, opteron, myrinet*
- E. g., `#PBS -l nodes=1:x86:nomyrinet`
- Or `#PBS -l nodes=32:x86:myrinet:ppn=2`
- Maximize resource pool
- Leave ***walltime*** and ***mem*** buffer (slide 22)

qsub Memory Specs

- **If single processor job and default memory suffices:**
  - Do not specify any memory settings
- **If single processor job needs > 400mb:**
  - E. g., `qsub -l mem=500mb`
- **If multi-processor job and default memory per processor suffices:**
  - Do not specify any memory settings
- **If multi-processor job needs > 400mb per processor, e. g., for 10 processors:**
  - Use *pmem=* and *mem=* options
  - E.g., `qsub -l pmem=500mb, mem=5000mb`

ACCRE Advanced Computing Center for Research & Education

Using The Scheduler

`qsub [options] <pbs_script>`	submit job for execution
`qstat`	view job(s) status
`showq`	view queue status
`pbsnodes -l -a`	view nodes & attributes
`checkjob -v <jobID(s)>`	view job(s) status
`checknode <nodename>`	view node status
`mdiag -f`	check fairshare
`mdiag -v -p`	check job priority
`mdiag -v -j <jobID>`	resource summary
`tracejob -n <#days> <jobID>`	trace job history

Self Diagnosing Problems

- Killed jobs
  - Bug in your code or script
  - Scheduler killed because exceeded resources, e. g., walltime, memory.
    - Leave a buffer in these parameters - especially with unfamiliar, new, or newly scaled-up code
    - Also, unexpected high system load can slow running
  - Use linux `pmap` on node to estimate memory usage of running job
  - Use `p_reaper` in your PBS script to auto-kill jobs that cause memory problems, see:

    accre-forum 2007 March archive

Self Diagnosing Problems

- Blocked or Deferred jobs, e. g., ***too_much_mem.pbs***
  - Use **`checkjob -v`** to see the reason
  - **`qstat -f`** gives similar information
  - Changing parameters, **`qdel`**, and resubmitting
  - Or **`qalter/mjobctl`**
- Jobs that do not return results
  - Use **`tracejob`** on ***vmpsched***, note non-zero ***Exit_status***

ACCRE Advanced Computing Center for Research & Education

Self Diagnosing Problems

- Long wait times: check cluster utilization, fairshare, and job priority, and refine resource request if possible
  - `mdiag -f` (older command called `diagnose`)
  - `mdiag -v -p`
  - Look at utilization charts on website, especially by processor type:

http://www.accre.vanderbilt.edu/utilization/index.php

Self Diagnosing Problems

- Slow execution may be due to load on node, load on local or shared file system, or high network loads
  - `pbsnodes`
  - Briefly log onto node and use Unix:

    `uptime`, `top`, or `ps`
  - Log onto ***`vmpsched`*** to see offline nodes report
  - Please report problem nodes or slow connectivity through RT

ACCRE Advanced Computing Center for Research & Education

Scheduler Etiquette

- Our goal is to provide fair use of the resources
  - 100% fair usage
  - Set number of CPUs becoming free every hour
- Stage large quantity job submissions (10 idle jobs allowed at a time)
- To maximize your use of the available resources
  - Start modestly - test new or unfamiliar code on test cluster first
  - ssh you@test[dd|opt|ppc]gw1.accre.vanderbilt.edu
  - Learn scheduler commands from man pages, online docs, ACCRE site:

http://www.accre.vanderbilt.edu/support/selfhelp/faq.php#moabcommands

Scheduler Etiquette

- TORQUE/PBS and Moab scheduler and job submission documentation at Cluster Resources:

http://www.clusterresources.com/pages/resources/documentation.php

- Help for specific commands:
  - Under TORQUE Resource Manager follow these links:
    - TORQUE Wiki Documentation
    - Documentation overview
    - A. Commands overview
  - Under Moab Workload Manager follow these links:
    - Commands Documentation

Scheduler Etiquette

- To maximize your use of the available resources (cont'd)
  - Know your code, available cluster resources vs. required resources
  - Know cluster policies on runtime and resource limitations (continually updated a `qsub` prefilter to catch runtime incompatibilities): http://www.accre.vanderbilt.edu/mission/cluster_policies
  - Plan ahead for long jobs
  - If possible, compile code on x86 & PPC architectures
  - Ask experienced group members (if possible)
  - Ask us (submit RT) if must run in unusual way

ACCRE Advanced Computing Center for Research & Education

Cluster Storage/Backup

- Cluster
  - GPFS file system from IBM

www.accre.vanderbilt.edu/mission/services/hpc.php#gpfs

  - can store your data on **`/home`** and **`/scratch`**
  - **`/home`** backed up daily using TiBS

www.accre.vanderbilt.edu/mission/services/storage.php

Cluster Storage/Backup

- Disk quotas
  - **`/home`** (10GB soft; 20GB hard)
  - **`/scratch`** (10GB soft; 100GB hard)
- File quotas
  - **`/home`** (100,000 soft; 200,000 hard)
  - **`/scratch`** (100,000 soft; 1,000,000 hard)
- GPFS **`mmlsquota`** shows your current total usage:

```
/usr/lpp/mmfs/bin/mmlsquota
```

- For convenience add to your PATH **`/usr/lpp/mmfs/bin`**
- Unix **`du`** shows disk usage in a given directory

ACCRE Advanced Computing Center for Research & Education

ACCRE Storage Policies

- Cluster disk usage and quota policies summary:

www.accre.vanderbilt.edu/mission/cluster_policies/diskspace_backups.php

- If you need to store larger quantities of data than the default allowance, ACCRE will work with you to arrange alternatives most suited to your needs, e.g., storage depots:

http://www.accre.vanderbilt.edu/mission/services/storage.php

ACCRE Advanced Computing Center for Research & Education

Getting Help

- Get help from experienced group members
- Join accre-forum and user's group
  - http://www.accre.vanderbilt.edu/support/lists.php
- Help from ACCRE
  - Materials on our website: User Support, FAQ, Cluster Policies
  - http://www.accre.vanderbilt.edu/support/contact/submit_RT.php
  - Office hours M-F 4-5PM