Running Jobs on WestGrid

Martin Siegert, SFU
Roman Baranowski, UBC
Stephen Cartwright, UofC
Masao Fujinaga, UofA

http://www.westgrid.ca/support/
Which is the right system for my program?
How do I submit a program?
How do I write a submission script?
What should I know about scheduling policies so that I can adjust my jobs for optimum efficiency?
How do I monitor submitted jobs?
Which System for which Job?

Cortex et al.: Parallel jobs using shared memory or MPI, large memory jobs, short run time

Nexus et al.: Parallel jobs using shared memory or MPI, large memory jobs, short run time

Matrix: Parallel jobs using MPI

Lattice: Parallel jobs using MPI, shared memory jobs up to 4 proc., Gaussian jobs

Glacier: Serial jobs, naturally parallel jobs using MPI, not more than 2GB of memory per process

Robson: Serial jobs, naturally parallel jobs using MPI, long run times
Submitting a Job

All WestGrid systems use a queuing and scheduling system Torque/Moab in order to:

- maximize the use of the resources
- distribute resources fairly between users

running a job:

- write a submission script (text editor)
- submit script using the qsub command
- results in \(<jobname>\).o<jobid>
- errors in \(<jobname>\).e<jobid>
Example

Submission script:
```
#!/bin/bash
#PBS -N hello-job
#PBS -l walltime=00:00:01
#PBS -l nodes=1

cd $PBS_O_WORKDIR
./hello
```

Job submission:
```
siegent@nunatak2:~/c-progs/example> qsub wg-serial.pbs
2837522.teva.westgrid.ubc
```

Jobid: 2837522

Errors:
```
siegent@nunatak2:~/c-progs/example> ls -l
```

Output:
```
siegent@nunatak2:~/c-progs/example> ls -l
-rw-r-xr-x 1 siegent siegent 4670 Jan 3 12:42 hello
-rw-r--r-- 1 siegent siegent 237 Jan 3 12:41 hello.c
-rw------- 1 siegent siegent 0 Jan 3 13:52 hello-job.e2837522
-rw------- 1 siegent siegent 13 Jan 3 13:52 hello-job.o2837522
-rw-r--r-- 1 siegent siegent 99 Jan 3 13:46 wg-serial.pbs
```

```
#!/bin/bash
#PBS -N wg-serial
#PBS -l walltime=48:00:00
#PBS -M siegert@sfu.ca
#PBS -m ae
#PBS -r n
#PBS -l nodes=1

cd $PBS_O_WORKDIR
./myprogram

Save script in file, e.g., wg-serial.pbs
Submit job with: qsub wg-serial.pbs
Parallel Job Scripts

Very system dependent!
(see system dependent section later)

Details on the web site:
www.westgrid.ca/support/programming/#parallel
Priority: Fairshare

Fairshare algorithm:

- every WestGrid project has an assigned fairshare target: usage percentage of the system
- priority depends on the difference between the target and the actual usage
- the actual usage is calculated within several fairshare windows with exponentially decreasing weights, i.e., historic usage is taken into account, but with decaying weights.
Monitoring

List of jobs:

- `qstat -a`
- `qstat -u <username>`
- `showq`
- `showq -r` (running)
- `showq -i` (idle)
- `showq -b` (blocked)

WG Portal https://portal.westgrid.ca/login.php
estimated start time:

showstart <jobid>
problems?
checkjob <jobid>

disappeared?
tracejob <jobid>
tracejob -n 4 <jobid>
Mainly parallel or large memory jobs
Serial jobs allowed on headnode
Short runtimes (24 hours)
IBM Power5 processors (plus one machine with Power4 processors)
4, 32, or 64 processor machines with 16, 156, or 256 GB memory, respectively
http://www.ualberta.ca/AICT/RESEARCH/WestGrid
Jobs for the Power5 machines are submitted to a single queue and job sent to a machine that satisfies the resource requirement requested.

Jobs for the Power4 machine are submitted to a different queue (pwr4).

Priorities according to fairshare.
MPI job on cortex

#!/usr/bin/bash
#PBS -S /usr/bin/bash
#PBS -l ncpus=2
cd $PBS_O_WORKDIR
# Note: MP_PROCS should be set to the number of processors required.
export MP_PROCS=2
export MP_HOSTFILE=/usr/local/etc/mpihosts/host.list
export MP_SHARED_MEMORY=yes
./pn
## Machine limits

<table>
<thead>
<tr>
<th>Machine name</th>
<th>Duration (hours)</th>
<th>Min. ncpus</th>
<th>Max. ncpus</th>
<th>Max. memory Gbytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>cortex</td>
<td>24</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>dendrite</td>
<td>24</td>
<td>3</td>
<td>64</td>
<td>256</td>
</tr>
<tr>
<td>synapse</td>
<td>24</td>
<td>32</td>
<td>62</td>
<td>256</td>
</tr>
<tr>
<td>bigfoot</td>
<td>24</td>
<td>1</td>
<td>32</td>
<td>16</td>
</tr>
</tbody>
</table>
Mainly parallel or large memory jobs
Serial jobs allowed on two of the machines
Short runtimes (24 hours)
SGI MIPS processors of varying speeds
8 to 256 processor machines with 8 to 256 GB memory
http://www.ualberta.ca/AICT/RESEARCH/WestGrid
Jobs are submitted to different queues corresponding to each machine.

Each queue has minimum and maximum limits on the number of CPUs as well as the amount of memory.

Priorities according to fairshare.

Job priority increases with time in queue.

Job priority decreases with recently run jobs.
MPI job on nexus

#!/usr/bin/bash
#PBS -S /usr/bin/bash
#PBS -q nexus
#PBS -l ncpus=2

# Script for running MPI sample program pn on nexus

cd $PBS_O_WORKDIR

# Note: NCPUS is a variable set by TORQUE to match the ncpus request above.

mpirun -np $NCPUS ./pn
## Queue limits

<table>
<thead>
<tr>
<th>queue name</th>
<th>duration (hours)</th>
<th>min ncpus</th>
<th>max available ncpus</th>
<th>max available memory (GBytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nexus</td>
<td>24</td>
<td>1</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>arcturus</td>
<td>24</td>
<td>64</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>borealis</td>
<td>24</td>
<td>8</td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>australis</td>
<td>24</td>
<td>16</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>corona</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>helios</td>
<td>24</td>
<td>1</td>
<td>32</td>
<td>16</td>
</tr>
</tbody>
</table>
4 CPU Alpha machines (667 MHz – 1250 Mhz)
Tru64 Unix operating system
17 DEC ES40s and 37 ES45s
ES45s have Quadrics interconnect
ES45s for parallel jobs, ideally 4-16 CPUs
ES40s only good for 1-4 CPU jobs
Memory available on a single machine ranges from 2GB to 8 GB
Most machines have 4 GB
Can be long wait times for 8 GB nodes
Lattice Resources

- 10GB quota for home
- /scratch and /scratch2 available for running jobs
  - 1.5 TB and 1.6 TB respectively
  - Not for long term storage!
- Most machines have local scratch (/local_scratch) with 100-136 GB capacity
  - This local disk has been found to be six times faster in a Gaussian test vs. non-local disk
- No quota on scratch filesystems, so please be reasonable with usage
  - Clean up old files!
Parallel jobs are preferred, however serial jobs are allowed to run.

G03 queue can be accessed using qsub's "-q" flag: i.e. "qsub -q g03"

Jobs submitted to this queue have access to local scratch and three weeks (503 hours) run time.

G03 queue sends jobs to ES40s and ES45s, default queue jobs only go to ES45s.

No jobs > 4 CPUS submitted to g03 queue.
Run time for default queue is restricted to one week (168 hours)

There are also two nodes that only allow short (3 hour) jobs and ten nodes that only allow 24 hour jobs

One interactive node (wg1) which is accessed using the interactive queue “qsub -q interactive” and limited to 1 hour run time
Matrix Overview

- 128 nodes
- Dual AMD 2.4 GHz Opteron processors
- 2 GB of memory per node
- Fast Infiniband interconnect, good for parallel jobs
- 832 GB space for home (all users)
- 3.1 TB /scratch
- No quota so please be reasonable
- Linux operating system
Very homogeneous environment
No serial jobs except for debugging and benchmarking
Great for larger (i.e. 16-64) CPU jobs that can make use of the fast interconnect
Max walltime is 3 days (72 hours)
Two nodes (m1 and m2) reserved for 3 hour jobs
Allows for testing
“tracejob” command is not available for users on Matrix
serial jobs, naturally parallel MPI jobs (GigE)

3 head-nodes - nunatak{1,2,3}.westgrid.ca
glacier.westgrid.ca

840 nodes x 2 Intel Xeon CPU 3.06GHz (32bit)

90% nodes 2GB; 1 rack ice55_1,...,ice60_14 4GB

http://guide.westgrid.ca/
Glacier Policies

- Maximum run time: 10 days
  
  ```bash
  #PBS -l walltime=240:00:00
  ```

  Job needs more time – send e-mail in advance

  Default runtime = 3h

- Nodes are shared ==> specify memory for job
  
  ```bash
  #PBS -l mem=1024mb
  ```

  Default memory = 768mb (memory violation policy in effect)

- Parallel job (more than 4 cpus) ask for “parallel”
  
  ```bash
  #PBS -l qos=parallel
  ```

  Quality of Service (qos)
Glacier Policies

- Short debugging job (max 4 cpus, 10 min runtime) **ask** for “debug” Quality of Service (qos)
  ```bash
  #PBS -l qos=debug
  ice1_1+ice1_2
  ```

- For a serial job maximum available memory
  - 2007MB on 2GB node
  - 4005MB on 4GB node

- Combine all “-l” options
  ```bash
  #PBS -l nodes=10,mem=10gb,qos=parallel
  ```
MATLAB users:

```
#PBS -W x=GRES:MATLAB,Image_Toolbox
```

If your job requires multiple software licenses

```
#PBS -W x=GRES:MATLAB+2
#PBS -W x=GRES:MATLAB+2,Image_Toolbox+2
```
Glacier job monitoring and statistics

```
qsort -c

Chassis   Load   Jobs  Down  Busy  Free  Down  Busy  Free
--------- --------- -------- -------- -------- -------- -------- -------- --------
iceflow1  18.96   24      0      12      2      0      24      4
iceflow10 20.9     28      0      14      0      0      28      0
iceflow50 14.54   28      0      14      0      0      28      0
iceflow51 17.97   28      0      14      0      0      28      0
iceflow52 15.54   28      0      14      0      0      28      0
iceflow53 17.74   28      0      14      0      0      28      0
iceflow54 18.79   28      0      14      0      0      28      0
iceflow55 26.7    28      0      14      0      0      28      0
iceflow56 25.14   28      0      14      0      0      28      0
iceflow57 21.62   28      0      14      0      0      28      0
iceflow58 25.64   28      0      14      0      0      28      0
iceflow59 25.51   28      0      14      0      0      28      0
iceflow60 21.11   26      1      13      0      2      26      0

Total: Unique jobs running = 558; Load = 1195.11
CPUs: Down = 2; Busy = 1671; Free = 7
Cluster utilization =~ 99.4%
```

```
/ /global/system/common/Stats/stats-Jan08
```
serial jobs, naturally parallel MPI jobs (GigE)
long runtimes (no limit for serial jobs)
54 PowerPC 970 processors (1.6 GHz, 64 bit)
2 proc./node, 4GB/node, 24GB swap/node
http://www.westgrid.ca/support/robson
max. 8 processors/user or 1 job using 8 to 24 processors

max. run time: $112 \text{ days/} (# \text{ of processors})$

e.g., 2 weeks for 8 processor job

no limit for serial jobs (must specify walltime in submission script)

no limits for preemptible jobs

no increase of job priority due to waiting time in queue
#!/bin/bash
# torque script for MPI job on robson
#PBS -N robson-par
#PBS -l walltime=168:00:00
#PBS -r n
#PBS -l nodes=16

cd $PBS_O_WORKDIR
mpiexec -machinefile $PBS_NODEFILE -n $PBS_NCPUS ./mpi-program

Do not use nodes=8:ppn=2

details at www.westgrid.ca/support/robson
Preemptible Jobs: the system is allowed to kill job when job with a higher priority gets submitted.

Checkpoint periodically: program has to write data to file that then can be used to restart job.

resubmit automatically: add to submission script

```
#PBS -r y
```

you decide:

- submit with: `qsub -q pre script.pbs`
- or add line to script file:
  
  ```
  #PBS -q pre
  ```

high throughput
Questions?