

ATLANTIC COMPUTATIONAL EXCELLENCE NETWORK

Parallel Programming Facilities at ACEnet

This tutorial

- How do I run...
 - OpenMP (multi-threaded) programs?
 - MPI (message-passing) jobs?
- What resources are available?
- How do I get them from the job scheduler?

What is ACEnet?

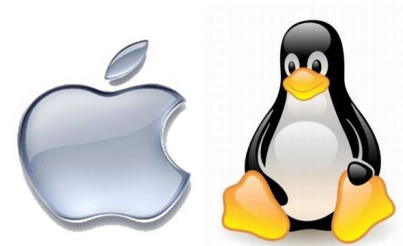
- Shared, publicly-funded resource...
- ...for *research computing*...
 - physics, chemistry, oceanography, biology, math, engineering, CS, ...
- ...across Atlantic Canada
 - four clusters, plus software
 - ~12 technical personnel
 - ~600 active accounts
- See <http://www.ace-net.ca/wiki/ACEnet>

- Common login for all clusters
- Separate filesystem for each cluster
 - Use sftp or scp to move data about
- All CPUs are **x86_64**
 - binaries are portable between clusters
 - CPU models, motherboards, other details vary
 - OS is Red Hat Enterprise Linux, various versions
- **Which cluster should I use?**
 - Try **glooscap.ace-net.ca**
 - but if it's busy, move anywhere

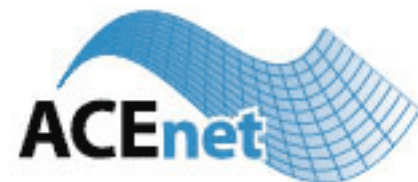
Getting Connected



- Logging in
 - `ssh` *Secure Shell*
- Moving Data Around
 - `sftp` *Secure File Transfer Protocol*
 - `scp` *Secure Copy*
- all bundled with **Mac OS X** or **Linux**
 - `ssh -X username@glooscap.ace-net.ca`
 - `sftp username@glooscap.ace-net.ca`
- **Windows?** Use PuTTY and WinSCP
or MobaXterm
just google for them



Command-Line Interface



```
glooscap-head.cs.dal.ca - PuTTY
Using username "rdickson".
rdickson@glooscap.ace-net.ca's password:
Last login: Wed Mar 16 10:03:16 2011 from rdickson.ucis.dal.ca
Welcome to Glooscap.ace-net.ca - Halifax, Nova Scotia
*****

rdickson@glooscap: ~ $ /usr/local/R/bin/R

R version 2.8.0 (2008-10-20)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> █
```

Environment modules

- Several compiler suites available
- and different MPI implementations
- **Advice:** Stick to the defaults
 - Portland Group (PGI) compilers, v.8.0-6
 - Open MPI v.1.2.9
- If not, use *modules*:

```
module avail; module help  
module load
```

Parallel programming APIs

- Shared-memory programming
 - OpenMP is a *standard*
 - Many compilers *implement* it
- Distributed-memory programming
 - MPI is a *standard*
 - *Implementation* at ACEnet is Open MPI
 - Don't be confused by similarity in names!

Today's sample files

```
$ tar xf ~rdickson/demo/OpenMP-demo.tar  
$ tar xf ~rdickson/demo/MPI-demo.tar
```

OpenMP compilation

- Most work done with compiler directives:
 - `!$omp directive` in Fortran
 - `#pragma omp directive` in C, C++
- Must use compiler flag
 - `-mp` PGI suite (default)
 - `-openmp` Intel suite
 - `-fopenmp` GCC

```
$ pgcc -mp hello.c -o hello
```

OpenMP execution

Thread count set with environment variable

```
$ ./hello
Hello World from thread 0
There are 1 threads

$ export OMP_NUM_THREADS=4
$ ./hello
Hello World from thread 0
Hello World from thread 2
Hello World from thread 1
Hello World from thread 3
There are 4 threads
```

MPI at ACEnet

- default module: Open MPI v.1.2.9
- wraps PGI compilers
 - “mpicc” is wrapper for pgcc
 - “mpic++” wraps pgCC
 - “mpif90” wraps pgf90
- **man mpicc; man pgcc**
- see http://www.ace-net.ca/wiki/Open_MPI

MPI 'Hello, world'

```
$ mpicc rank.c -o rank
$ mpirun -np 2 rank
Hello from process 1
...and hello from rank 0.
$ mpirun -np 4 rank
Hello from process 2
Hello from process 3
Hello from process 1
...and hello from rank 0.
```

TotalView debugger

- Compile with `-g` to get symbols
 - same as any other debugger
- Ensure X11 working (try `xclock`)
 - did you
`ssh -X yourid@mahone.ace-net.ca ?`
- Detailed instructions at
<http://www.ace-net.ca/wiki/TotalView>

Running TotalView

on the head node

- OpenMP:

```
$ pgcc -mp -g prog.c -o prog  
$ export OMP_NUM_THREADS=2  
$ totalview ./prog
```

- MPI:

```
$ mpicc -g prog.c -o prog  
$ mpirun --debug -np 2 prog
```

Head node or compute node?

- ACEnet rule:
Don't run tasks on login node longer than 10 CPU minutes...
- **...or that will cause memory to swap**
 - use `top` to check mem use on head node
- Longer or larger tasks must go to compute nodes
- Use dynamic resource manager "Grid Engine"

Running TotalView

on compute nodes

- OpenMP:

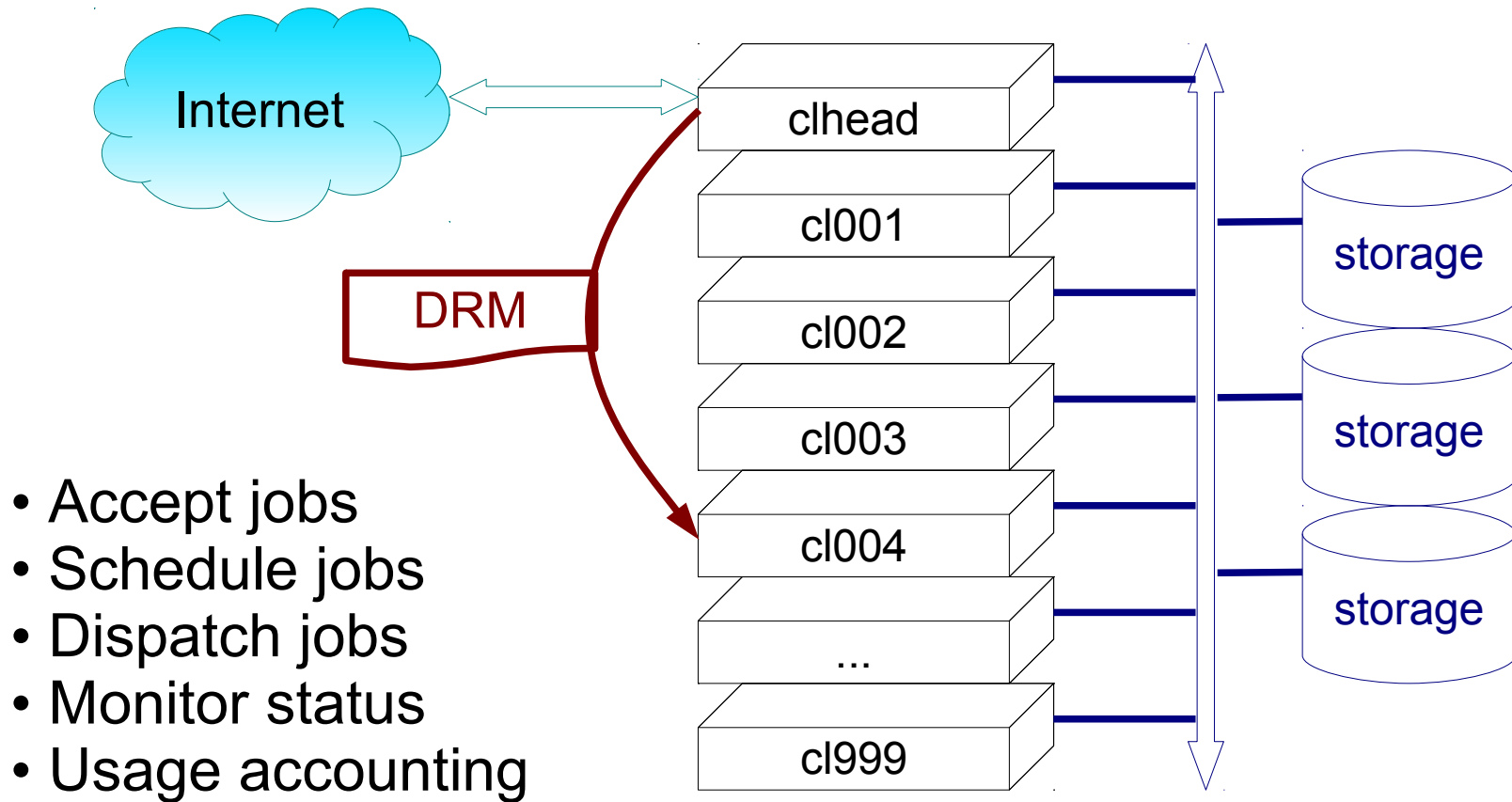
```
$ qcrsh -cwd -pe openmp 4 \  
-l h_rt=1:0:0,test=true \  
totalview ./prog
```

- MPI:

```
$ qcrsh -cwd -pe 'ompi*' 20 \  
-l h_rt=1:0:0,test=true \  
mpirun --debug prog
```

In other words, check the Wiki page...

Role of Dynamic Resource Manager



Dynamic Resource Managers



- Grid Engine used at ACEnet
 - *aka* Sun Grid Engine (SGE), N1GE, or Oracle Grid Engine
- Other DRMs and/or schedulers you might encounter:
 - OpenPBS, Torque, Moab, Maui, LoadLeveler, LSF, SLURM, ...
 - sometimes scheduler is distinct, sometimes bundled

Grid Engine takes a script



```
$ cat job.sh
```

```
## -cwd
```

```
## -l h_rt=0:10:0,test=true
```

```
## -pe omp* 4
```

```
mpirun hello
```

```
$ qsub job.sh
```

```
Your job 654321 ("job.sh") has been  
submitted
```

```
$ qstat
```

```
...
```

Grid Engine commands

- **qsub** [*options*] *script*
 - submit a job
- **qstat**
 - is my job running? done?
 - job vanishes from **qstat** when finished
- **qsum**^{*}
 - how busy is the cluster?
- **showq**^{*}
 - how far am I from the head of the waiting list?

^{*} Local scripts, no man pages, no options --- see www.ace-net.ca

qsub flags $\approx$ script directives



- Directives in script...

```
#$ -cwd  
#$ -l h_rt=1:0:0  
./myprog <file.in >file.out
```

- ...equivalent to arguments to qsub

```
qsub -cwd -l h_rt=1:0:0 script
```

- Command line trumps script
- Right-most (last) takes precedence

Some SGE parameters

- **-l h_rt=*hh:mm:ss***
 - hard limit on run time
 - **REQUIRED** - allows scheduler to plan ahead
- **-l h_vmem=*memory***
 - memory *per parallel slot*
- **-cwd**
 - run in current working directory
- **-j yes**
 - join stderr with stdout

Parallel environments

- `-pe parallel_env n_procs`
 - `parallel_env` one of
 - `mpi*`
 - `openmp`
 - `1per, 2per, 4per, 16per`
 - `n_procs` can be
 - a number,
 - a range, e.g. 4-8
 - or a list, e.g. 4,8,12,16

Local scratch disk

- Heterogeneous hardware
- It's complicated... see http://www.ace-net.ca/wiki/Local_Scratch
- For multi-host work, you may use `/scratch/tmp/$USER`, but be sure you clean up after yourself
 - especially failed jobs!
 - get host list from `$PE_HOSTFILE`

Some more Grid Engine commands



- **qdel *jobid***
- **qacct -j *jobid***
 - accounting record, may contain failure hints
- **qhost**
 - host statuses, memory, *etc.*
- **qrsh**
 - interactive session on compute host(s)
- **memslots *memory_amount***
 - See if memory is limiting resource*

* Local script, no man pages, no options --- see www.ace-net.ca

Where to go for help

- ACEnet documentation
 - <http://www.ace-net.ca/wiki/ACEnet>
 - *'search' is your friend*
- Systems status, *etc.*
 - http://www.ace-net.ca/wiki/Cluster_Status
- User support
 - email: support@ace-net.ca
 - see guidelines at http://www.ace-net.ca/wiki/Ask_Support