Cluster Workshop

UCDAVIS genome center

Introduction

"If you were plowing a field, which would you rather use: Two strong oxen or 1024 chickens?"

–Seymour Cray

Introduction

- What is a cluster?

Introduction

- What is a cluster?
  - ⇒ A collection of machines that work together

Introduction

- What is a cluster?
  - ⇒ A collection of machines that work together
- Why use a cluster?

Introduction

- What is a cluster?
  - ⇒ A collection of machines that work together
- Why use a cluster?
  - ⇒ Chickens are cheaper than oxen

Introduction

- What is a cluster?
  - ⇒ A collection of machines that work together
- Why use a cluster?
  - ⇒ Chickens are cheaper than oxen
  - ⇒ (and easier to feed)

Introduction

- What is a cluster?
  - ⇒ A collection of machines that work together
- Why use a cluster?
  - ⇒ Chickens are cheaper than oxen
  - ⇒ (and easier to feed)
  - ⇒ ...but try making 1024 chickens move in the same direction

Types of Jobs

- **interactive** A single-part job to be run immediately.

Types of Jobs

- **interactive** A single-part job to be run immediately.
- **batch (serial)** A single-part job to run in the background

Types of Jobs

- **interactive** A single-part job to be run immediately.
- **batch (serial)** A single-part job to run in the background
- **parallel** A job that has been split into multiple parts to run on more than one processor.

Types of Jobs

Parallel cluster jobs can be categorized according to the amount of communication required between parts of the job:

- **fine-grained parallel** Substantial communication required.

Types of Jobs

Parallel cluster jobs can be categorized according to the amount of communication required between parts of the job:

- **fine-grained parallel** Substantial communication required.
- **course-grained parallel** Occasional communication required.

Types of Jobs

Parallel cluster jobs can be categorized according to the amount of communication required between parts of the job:

- **fine-grained parallel** Substantial communication required.
- **course-grained parallel** Occasional communication required.
- **embarrassingly parallel** Almost no communication required.

Types of Jobs

Even serial jobs that cannot be split up to run in parallel can still benefit from a cluster environment, for example, by running with different sets of input files or parameters simultaneously.

Cluster Terminology

- **Node** An individual computer in the cluster

Cluster Terminology

- **Node** An individual computer in the cluster
- **Head Node** The main node. Coordinates scheduling jobs among the nodes. This is what you log in to from outside.

Cluster Terminology

- **Node** An individual computer in the cluster
- **Head Node** The main node. Coordinates scheduling jobs among the nodes. This is what you log in to from outside.
- **Interconnect** The network or networks that connects the nodes together

Cluster Terminology

- **Node** An individual computer in the cluster
- **Head Node** The main node. Coordinates scheduling jobs among the nodes. This is what you log in to from outside.
- **Interconnect** The network or networks that connects the nodes together
- **MPI M**essage **P**assing **I**nterface, a protocol used for communication between parts of a job.

Cluster Storage

- Home directory

Cluster Storage

- Home directory
- Shared cluster space

Cluster Storage

- Home directory
- Shared cluster space
- Local storage

Storage: Home directory

- Network mounted

Storage: Home directory

- Network mounted
- Geared towards wide availability, not high performance

Storage: Home directory

- Network mounted
- Geared towards wide availability, not high performance
- Good place for your code, final results

Storage: shared

- Network shared, but only within the cluster

Storage: shared

- Network shared, but only within the cluster
- Medium availability, medium performance

Storage: shared

- Network shared, but only within the cluster
- Medium availability, medium performance
- Good place for dataset that all nodes need to access

Storage: shared

- Network shared, but only within the cluster
- Medium availability, medium performance
- Good place for dataset that all nodes need to access
- Performance can suffer with many simultaneous accesses

Storage: local

- Not network shared, only available to a single node

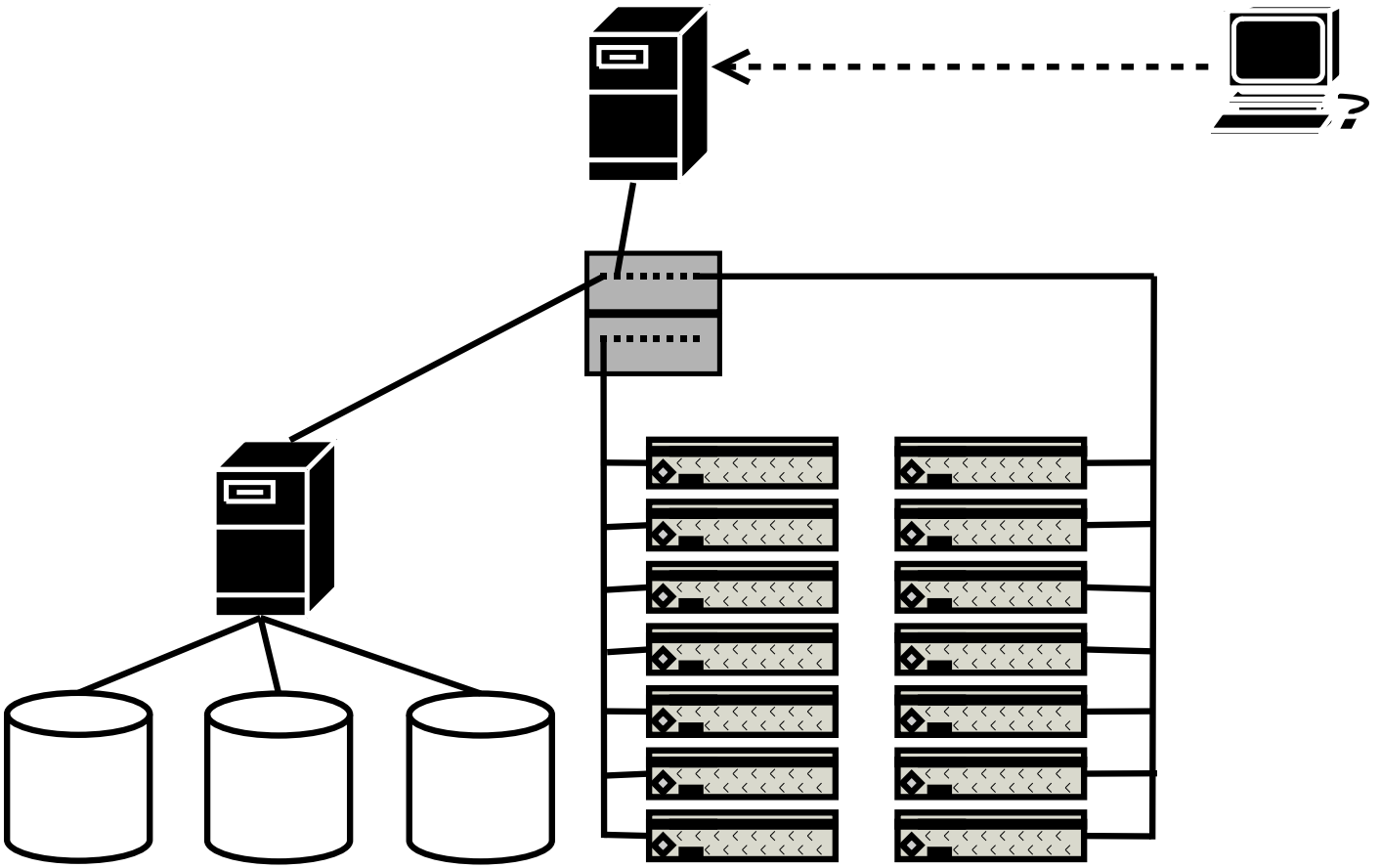
Storage: local

- Not network shared, only available to a single node
- Highest performance, but least convenient

Storage: local

- Not network shared, only available to a single node
- Highest performance, but least convenient
- Good place for working set

Anatomy of a Cluster



Accessing the cluster

- Access using ssh

Accessing the cluster

- Access using ssh
- Transfer files using scp/sftp/rsync

Accessing the cluster

- Access using ssh
- Transfer files using scp/sftp/rsync
- Web interface

Data Transfer

- scp

Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`

Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync

Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync
  - ⇒ `rsync -a directory username@genbeo:`

Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync
  - ⇒ `rsync -a directory username@genbeo:`
- sftp

Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync
  - ⇒ `rsync -a directory username@genbeo:`
- sftp
  - ⇒ ftp-like interface that works over ssh

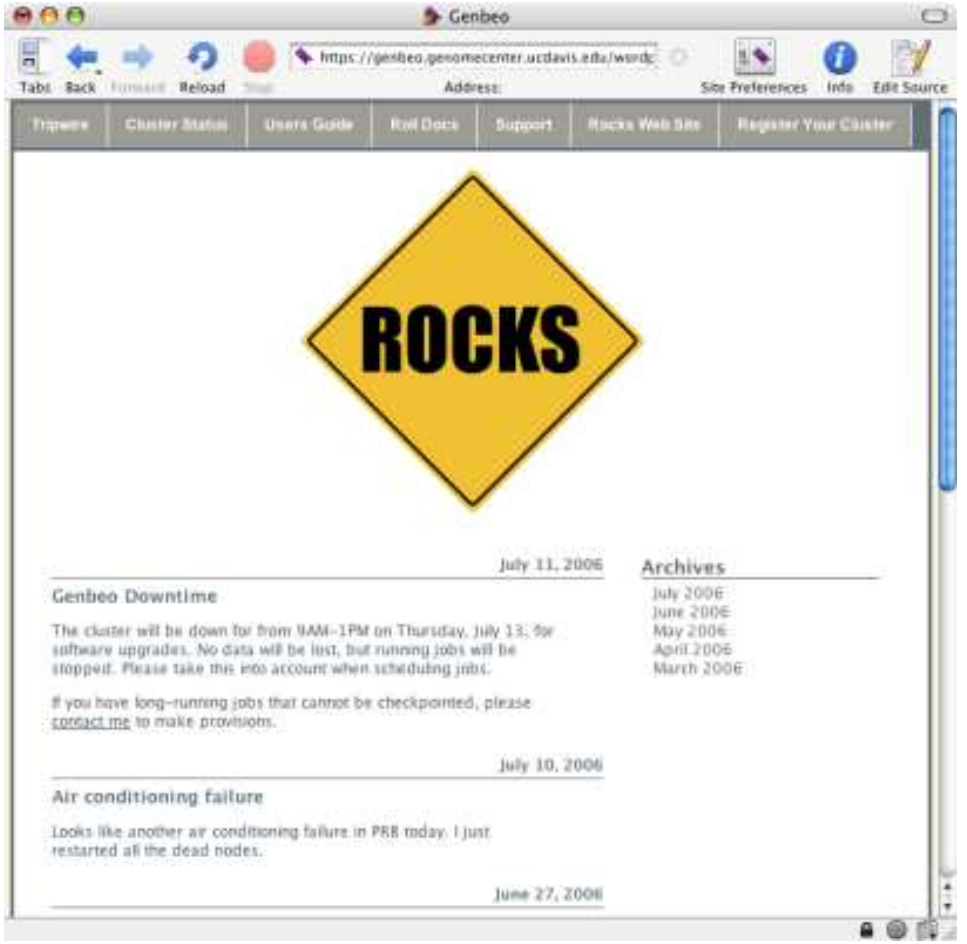
Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync
  - ⇒ `rsync -a directory username@genbeo:`
- sftp
  - ⇒ ftp-like interface that works over ssh
- wget

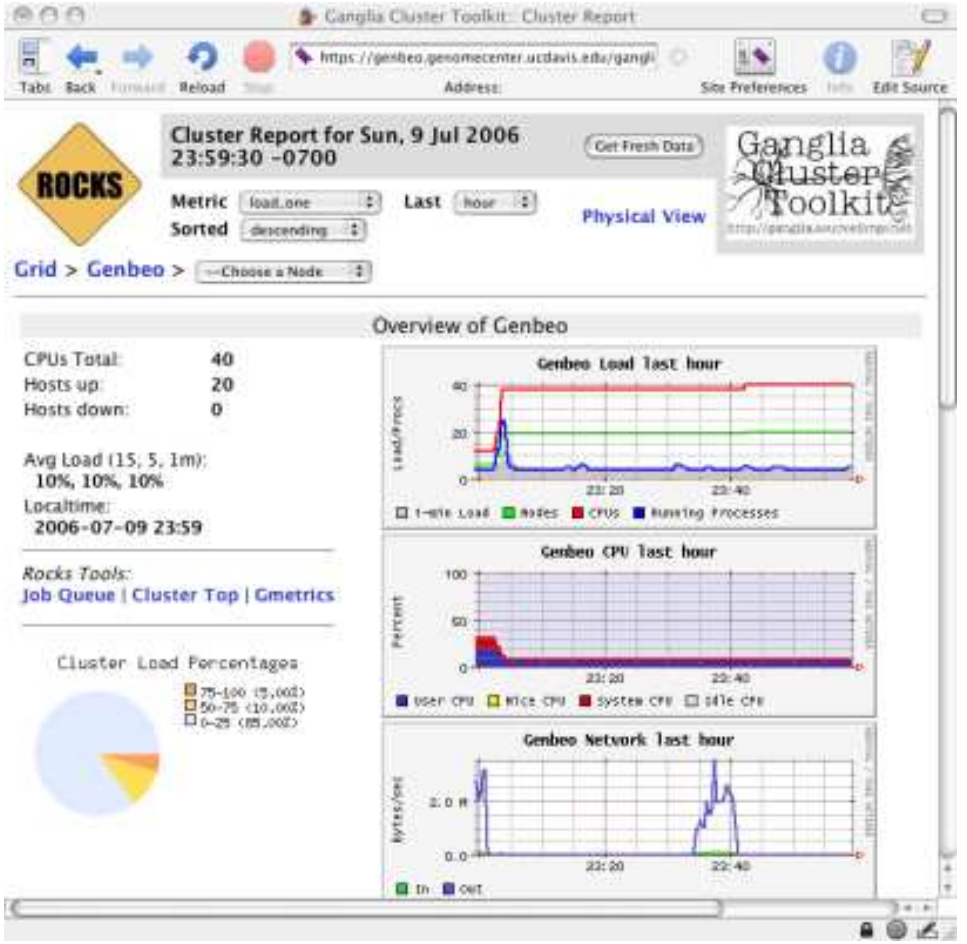
Data Transfer

- scp
  - ⇒ `scp file username@genbeo:`
- rsync
  - ⇒ `rsync -a directory username@genbeo:`
- sftp
  - ⇒ ftp-like interface that works over ssh
- wget
  - ⇒ client to download files from the web directly to the cluster

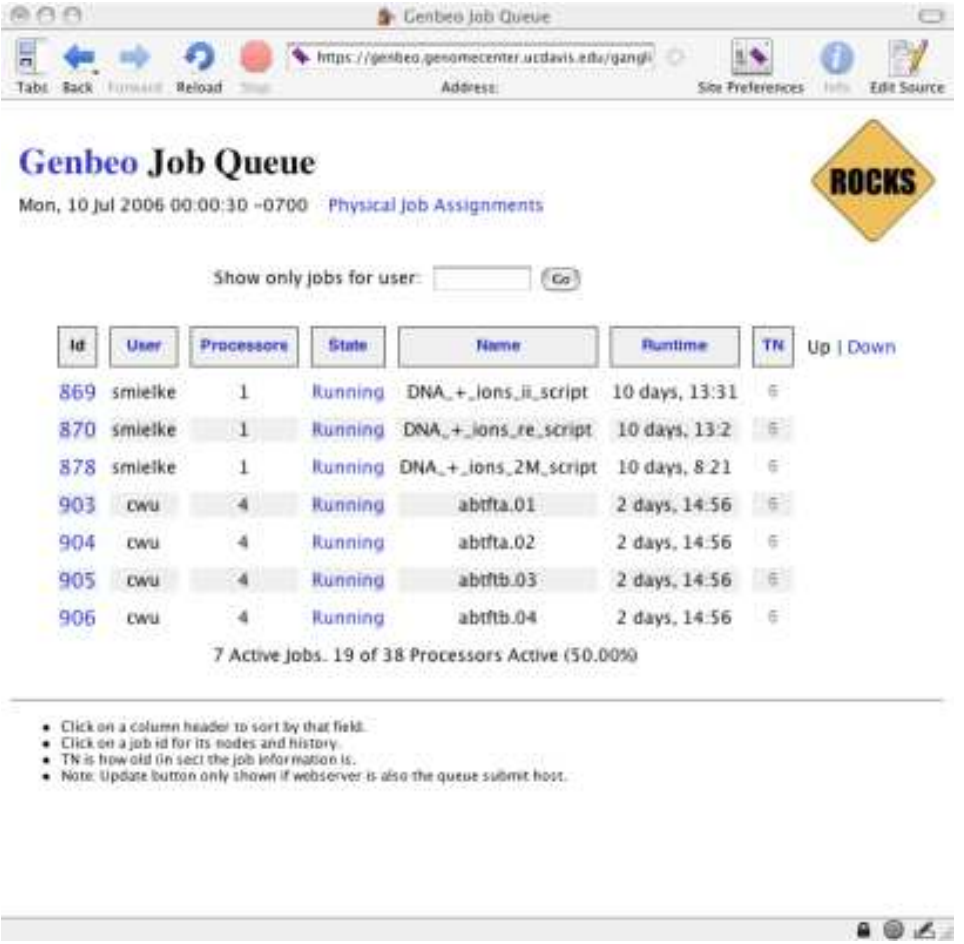
Web Interface



Cluster Web Interface



Cluster Web Interface



Id	User	Processors	State	Name	Runtime	TN
869	smielke	1	Running	DNA_+_ions_ii_script	10 days, 13:31	6
870	smielke	1	Running	DNA_+_ions_re_script	10 days, 13:2	6
878	smielke	1	Running	DNA_+_ions_2M_script	10 days, 8:21	6
903	cwu	4	Running	abtfta.01	2 days, 14:56	6
904	cwu	4	Running	abtfta.02	2 days, 14:56	6
905	cwu	4	Running	abtftb.03	2 days, 14:56	6
906	cwu	4	Running	abtftb.04	2 days, 14:56	6

Sun Grid Engine

- What is a batch queue?

Sun Grid Engine

- What is a batch queue?
  - ⇒ Manage cluster resources

Sun Grid Engine

- What is a batch queue?
  - ⇒ Manage cluster resources
- Why use the batch queue?

Sun Grid Engine

- What is a batch queue?
  - ⇒ Manage cluster resources
- Why use the batch queue?
  - ⇒ Share cluster resources

Sun Grid Engine

- What is a batch queue?
  - ⇒ Manage cluster resources
- Why use the batch queue?
  - ⇒ Share cluster resources
  - ⇒ You don’t need to worry about when/where your jobs run

Sun Grid Engine

- What is a batch queue?
  - ⇒ Manage cluster resources
- Why use the batch queue?
  - ⇒ Share cluster resources
  - ⇒ You don’t need to worry about when/where your jobs run
  - ⇒ Submit a whole bunch of jobs and go home!

The Cluster Café

You can think of the cluster as a restaurant:

- **Jobs** are parties coming to eat

The Cluster Café

You can think of the cluster as a restaurant:

- **Jobs** are parties coming to eat
- **Tables** are the nodes of the cluster

The Cluster Café

You can think of the cluster as a restaurant:

- **Jobs** are parties coming to eat
- **Tables** are the nodes of the cluster
- Usually the scheduler will try to place each job at its own table, but if it's a busy day, you might have to share a table with someone you don't know.

Slots

A **slot** is a placeholder where a job or part of a job can run. In the Cluster Café, it is a chair at a table.

- A resource allocated to your job.

Slots

A **slot** is a placeholder where a job or part of a job can run. In the Cluster Café, it is a chair at a table.

- A resource allocated to your job.
- We define one slot per CPU core

Slots

A **slot** is a placeholder where a job or part of a job can run. In the Cluster Café, it is a chair at a table.

- A resource allocated to your job.
- We define one slot per CPU core
- Request number of slots when you submit job

Slots

A **slot** is a placeholder where a job or part of a job can run. In the Cluster Café, it is a chair at a table.

- A resource allocated to your job.
- We define one slot per CPU core
- Request number of slots when you submit job
- Allocating a slot is only advisory: the scheduler has reserved the requested number of slots.

Resources

It is possible to request other resources when you submit your job

Resources

It is possible to request other resources when you submit your job

$\Rightarrow$"I'd like a table by the window"

Resources

It is possible to request other resources when you submit your job

$\Rightarrow$"I'd like a table by the window"

Or even to request a specific node

Resources

It is possible to request other resources when you submit your job

⇒"I'd like a table by the window"

Or even to request a specific node

⇒"I want to sit at table 2. I'll wait."

Resources

It is possible to request other resources when you submit your job

⇒"I'd like a table by the window"

Or even to request a specific node

⇒"I want to sit at table 2. I'll wait."

The scheduler will wait to run your job until it can fulfill your requirements.

Parallel Environment

A **Parallel Environment** sets up the resources required to run a multi-node job.

- Need to use a PE when you want more than one slot

Parallel Environment

A **Parallel Environment** sets up the resources required to run a multi-node job.

- Need to use a PE when you want more than one slot
- Request desired PE when you submit your job

Parallel Environment

A **Parallel Environment** sets up the resources required to run a multi-node job.

- Need to use a PE when you want more than one slot
- Request desired PE when you submit your job

Tells the scheduler how you'd like the table set

Parallel Environment

Available PEs:

- mpi Sets up environment for LAM MPI

Parallel Environment

Available PEs:

- mpi Sets up environment for LAM MPI
- mpich Sets up environment for mpich

Parallel Environment

Available PEs:

- mpi Sets up environment for LAM MPI
- mpich Sets up environment for mpich
- serial/threaded Makes sure all slots are on the same node, does not set up any inter-node communication environment.

Job Arrays

- Run the same job multiple times

Job Arrays

- Run the same job multiple times
- submit/manage as a single job

Job Arrays

- Run the same job multiple times
- submit/manage as a single job
- Ideal for running the same program repeatedly with different input files or parameters

Queue example

Three jobs:

- A parallel job using 6 slots
- A job array of 3 jobs, using one slot each
- A parallel job using 4 slots

Queue example

Job	Slots
A	6
B_1	1
B_2	1
B_3	1
C	4

Queue

Node1

Node2

Three jobs waiting in queue...

Queue example

Job	Slots
A	6
B_1	1
B_2	1
B_3	1
C	4

Queue

A	A
A	A

Node1

A	A

Node2

Job A is scheduled

Queue example

Job	Slots
A	6
B_1	1
B_2	1
B_3	1
C	4

Queue

A	A
A	A

Node1

A	A
B_1	

Node2

Job B_1 is scheduled

Queue example

Job	Slots
A	6
B_1	1
B_2	1
B_3	1
C	4

Queue

A	A
A	A

Node1

A	A
B_1	B_2

Node2

Job B_2 is scheduled

Queue example

Job	Slots
B_1	1
B_2	1
B_3	1
C	4

Queue

Node1

B_1	B_2

Node2

Job A finishes

Queue example

Job	Slots
B_1	1
B_2	1
B_3	1
C	4

Queue

	B_3

Node1

B_1	B_2

Node2

Job B_3 is scheduled

Queue example

Job	Slots
B_2	1
B_3	1
C	4

Queue

	B_3

Node1

	B_2

Node2

Job B_1 finishes

Queue example

Job	Slots
B_3	1
C	4

Queue

	B_3

Node1

Node2

Job B_2 finishes

Queue example

Job	Slots
B_3	1
C	4

Queue

	B_3

Node1

C	C
C	C

Node2

Job C is scheduled

SGE Commands

- qsub: submit jobs

SGE Commands

- qsub: submit jobs
- qstat: get job status

SGE Commands

- qsub: submit jobs
- qstat: get job status
- qdel: remove a job

SGE Commands

- qsub: submit jobs
- qstat: get job status
- qdel: remove a job
- qlogin: interactive login

SGE Commands: qsub

Use the `qsub` command to submit a batch job to the system

Simplest case:

```
$ qsub file.sh
Your job 929 ("file.sh") has been submitted.
```

SGE Commands: qstat

Use the `-f` flag to see all jobs running...

```
$ qstat -f
queuename                      qtype used/tot. load_avg arch        states
---------------------------------------------------------------------------
all.q@compute-0-1.local        BIP   4/4       1.83     lx26-amd64
  2455 0.60500 proAwt     cwu         r     06/21/2006 12:06:06   4
---------------------------------------------------------------------------
all.q@compute-0-10.local       BIP   0/4       0.00     lx26-amd64  d
---------------------------------------------------------------------------
all.q@compute-0-98.local       BIP   4/4       2.80     lx26-amd64
  2823 0.51386 ccr5_SCH_A twang       r     07/07/2006 15:12:51   2
  2865 0.50500 rungb5b    xjdeng      r     07/08/2006 17:37:06   1
  2944 0.52905 g2l_ff03   zxwang      r     07/09/2006 22:07:21   1
---------------------------------------------------------------------------
all.q@compute-0-99.local       BIP   0/4       0.00     lx26-amd64
```

SGE Commands: qstat

...as well as those waiting to be run.

```
#########################################################################
PENDING JOBS - PENDING JOBS - PENDING JOBS - PENDING JOBS - PENDING JOBS
#########################################################################
2947 0.52905 g2f_ff03   zxwang       qw    07/09/2006 22:11:04    20
2948 0.52905 g2h_ff03   zxwang       qw    07/09/2006 22:11:55    20
2949 0.52905 g2q_ff03   zxwang       qw    07/09/2006 22:12:42    20
```

SGE Commands: qstat

Use the `-j <jobid>` flag to get more information about a job:

```
$ qstat -j 2947
job_number:                 2947
exec_file:                  job_scripts/2947
submission_time:            Sun Jul  9 22:11:04 2006
...
```

SGE Commands: qdel

- Use the `qdel` command to delete a previously scheduled job from the queue.

SGE Commands: qdel

- Use the `qdel` command to delete a previously scheduled job from the queue.
- Note: if the job was running, you may still have to kill the processes by hand.

SGE Commands: qdel

- Use the `qdel` command to delete a previously scheduled job from the queue.
- Note: if the job was running, you may still have to kill the processes by hand.
- The `-f` (force) option can sometimes be necessary to clean up jobs left behind, for example, if a node dies during the job.

SGE Commands: qlogin

Use the `qlogin` command to schedule an interactive login.

- Default is one slot

SGE Commands: qlogin

Use the `qlogin` command to schedule an interactive login.

- Default is one slot
- To allocate more slots, use the parallel environment serial and the number of slots
  `qlogin -pe serial 2`

SGE Commands: qlogin

Use the `qlogin` command to schedule an interactive login.

- Default is one slot
- To allocate more slots, use the parallel environment serial and the number of slots
  `qlogin -pe serial 2`
- Two slots on genbeo will give you the whole node (4 on shiraz)

SGE Commands: qlogin

Use the `qlogin` command to schedule an interactive login.

- Default is one slot
- To allocate more slots, use the parallel environment serial and the number of slots
  `qlogin -pe serial 2`
- Two slots on genbeo will give you the whole node (4 on shiraz)
- If enough slots are not available, `qlogin` will fail.

Things to do

- Use the scheduler!

Things to do

- Use the scheduler!
- Checkpoint your job

Things to do

- Use the scheduler!
- Checkpoint your job
- Make use of local storage on the nodes for intermediate results

Things NOT to do

- Run jobs on the head node

Things NOT to do

- Run jobs on the head node
- Many simultaneous writes to network filesystem

Things NOT to do

- Run jobs on the head node
- Many simultaneous writes to network filesystem
- Go around scheduler and run directly on the nodes