HPC User Environment, Job Management with PBS/Loadleveler

Alexander B. Pacheco

User Services Consultant
LSU HPC & LONI
sys-help@loni.org

HPC Training Fall 2012
Louisiana State University
Baton Rouge
September 12, 2012

- Two major architectures.

Linux Clusters

- Vendor: Dell
- OS: Red Hat
- CPU: Intel Xeon

AIX Clusters

- Vendor: IBM
- OS: AIX
- CPU: Power 7

Linux Clusters

	Name	Peak TeraFLOPS/s	Location	Status	Login
LONI	QueenBee	50.7	ISB	Production	LONI
	Eric	4.7	LSU	Production	LONI
	Louie	4.7	Tulane	Production	LONI
	Oliver	4.7	ULL	Production	LONI
	Painter	4.7	LaTech	Production	LONI
	Poseidon	4.7	UNO	Production	LONI
	Satellite	4.7	Southern	Unknown	LONI
LSU HPC	Tezpur	15.3	LSU	Production	HPC
	Philip	3.5	LSU	Production	HPC
	SuperMike	146 (CPU) 66 (GPU)	LSU	Coming Soon	HPC

AIX Clusters

	Name	Peak TF/s	Location	Status	Login
LSU HPC	Pandora	6.8	LSU	Production	HPC

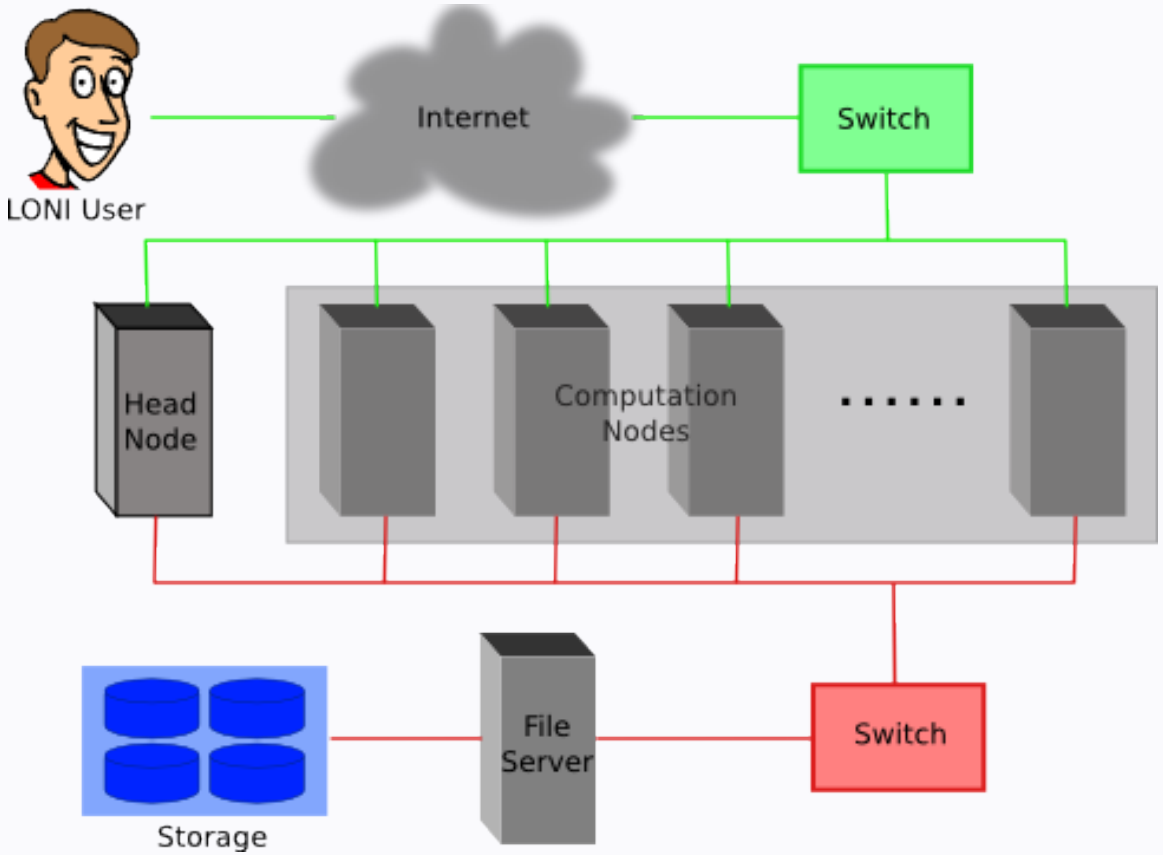
- LONI account

  `https://allocations.loni.org`
- LSU HPC account

  `https://accounts.hpc.lsu.edu`
- Newest cluster in production at LSU HPC is Pandora.
- Newest cluster at LSU HPC is SuperMike, estimated to be online sometime in Fall 2012.

- The default Login shell is bash
- Supported Shells: bash, tcsh, ksh, csh & sh
- Change Login Shell at the profile page

- A cluster is a group of computers (nodes) that works together closely
- Type of nodes
  - Head node
  - Compute node

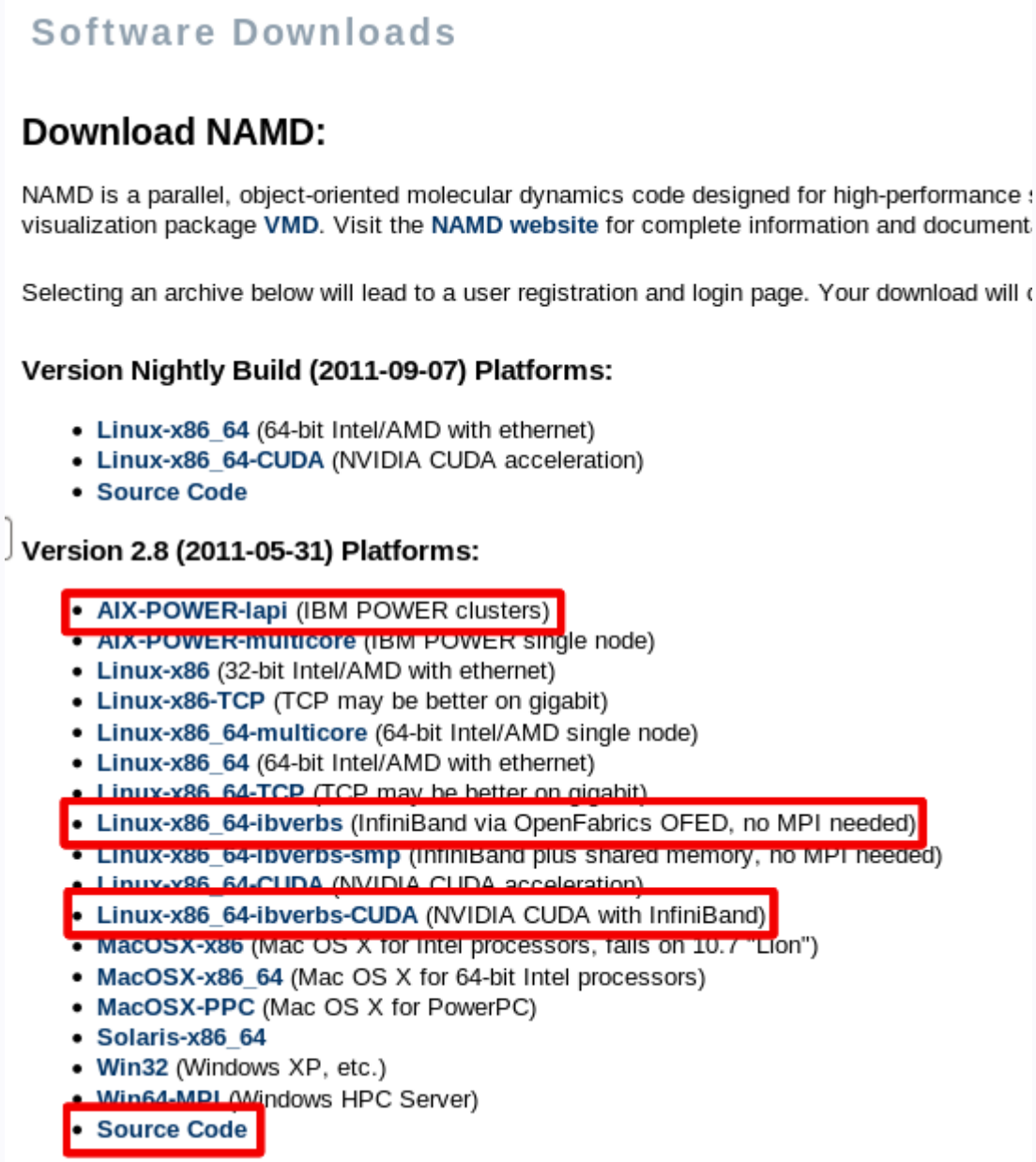


- Queen Bee
  - 668 nodes: 8 Intel Xeon cores @ 2.33 GHz
  - 8 GB RAM
  - 192 TB storage
- Other LONI clusters
  - 128 nodes: 4 Intel Xeons cores @ 2.33 GHz
  - 4 GB RAM
  - 9 TB storage
- Tezpur
  - 360 nodes, 4 Intel Xeon cores @ 2.33 GHz
  - 4 GB RAM
  - 32 TB storage
- Philip
  - 37 nodes, 8 Intel Xeon cores @ 2.93 GHz
  - 24/48/96 GB RAM
  - Shares storage with Tezpur
- Pandora
  - 8 Power7 nodes, 8 IBM Power7 processors @ 3.33 GHz
  - 128 GB RAM
  - 19 TB storage

- 2 Philip nodes with 3 GPU's each.
  - Intel Xeon X5650
    - 2.66GHz Dual Hexa-core with hyperthreading: 12 cores, 24 threads
    - Memory : 48GB
    - L3 Cache: 12MB
  - Tesla M2070
    - 448 CUDA cores (14 Multiprocessor $\times$ 32 CUDA Cores/MP)
    - GPU Clock Speed: 1.15GHz
    - Total Memory: 5.25GB
- Compilers: CUDA 4.0, PGI Accelerator
- Infiniband Interconnect between the two nodes.

- 146 CPU TFlops and 66 double-precision GPU TFlops,
- 382 standard nodes, dual Intel Sandy Bridge CPUs with 32GB RAM (16 cores per node),
- 50 GPU nodes, dual Intel Sandy Bridge CPUs with 64GB RAM and dual NVIDIA Tesla M2090 6GB GPUs,
- 8 big memory nodes with 256GB RAM, capable of aggregation into a single virtual symmetric processing (vSMP) node using ScaleMP,
- Mellanox Infiniband QDR network of 2:1 over-subscription.

- There are numerous different architectures in the HPC world.
- Choose the software to install or use depending on cluster architecture.
  1. Linux: EM64T, AMD64, X86_64
  2. AIX: Power7



Software Downloads

Download NAMD:

NAMD is a parallel, object-oriented molecular dynamics code designed for high-performance s
visualization package **VMD**. Visit the **NAMD website** for complete information and document

Selecting an archive below will lead to a user registration and login page. Your download will

Version Nightly Build (2011-09-07) Platforms:

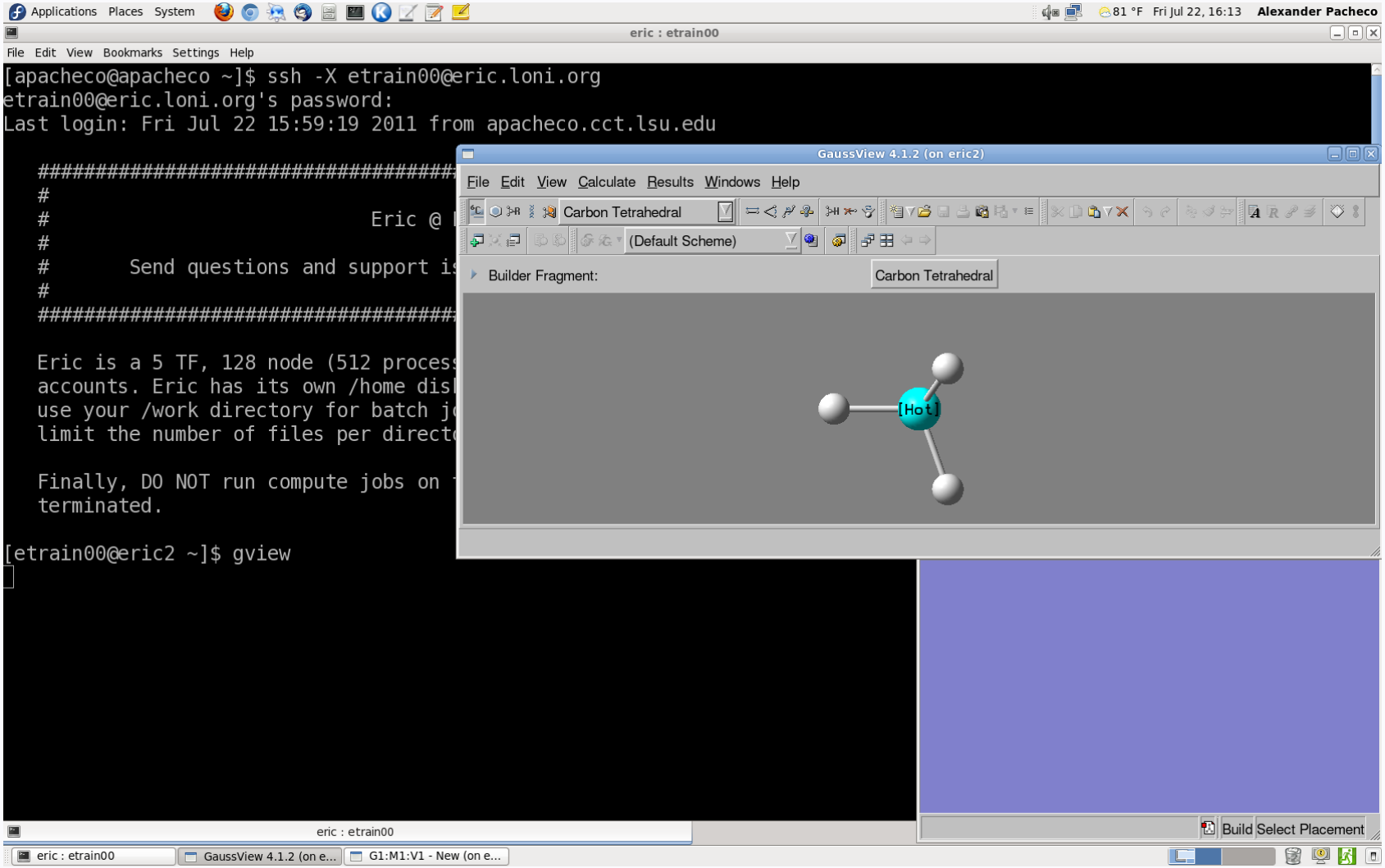
- **Linux-x86_64** (64-bit Intel/AMD with ethernet)
- **Linux-x86_64-CUDA** (NVIDIA CUDA acceleration)
- **Source Code**

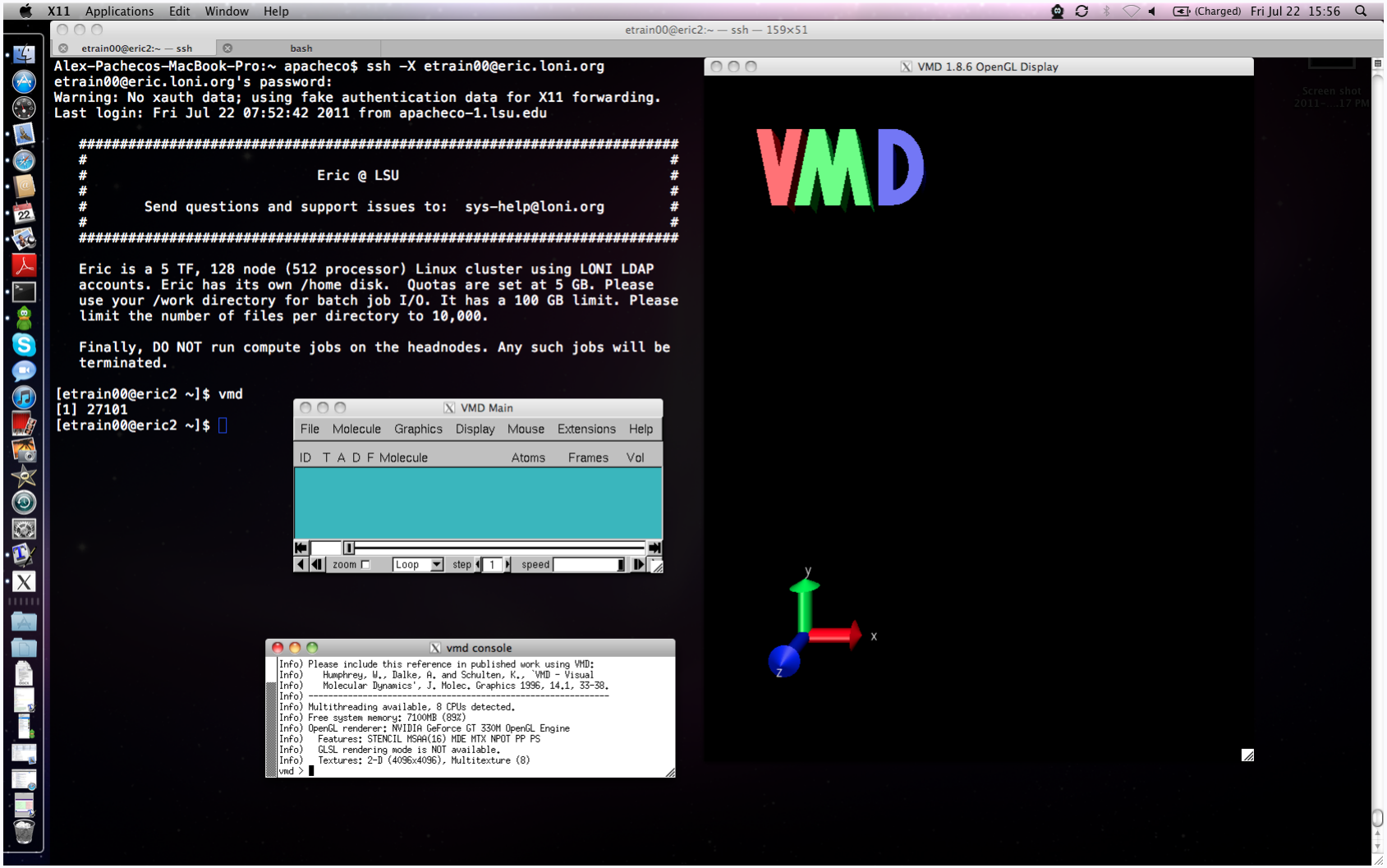
Version 2.8 (2011-05-31) Platforms:

- **AIX-POWER-lapi** (IBM POWER clusters)
- **AIX-POWER-multicore** (IBM POWER single node)
- **Linux-x86** (32-bit Intel/AMD with ethernet)
- **Linux-x86-TCP** (TCP may be better on gigabit)
- **Linux-x86_64-multicore** (64-bit Intel/AMD single node)
- **Linux-x86_64** (64-bit Intel/AMD with ethernet)
- **Linux-x86_64-TCP** (TCP may be better on gigabit)
- **Linux-x86_64-ibverbs** (InfiniBand via OpenFabrics OFED, no MPI needed)
- **Linux-x86_64-ibverbs-smp** (InfiniBand plus shared memory, no MPI needed)
- **Linux-x86_64-CUDA** (NVIDIA CUDA acceleration)
- **Linux-x86_64-ibverbs-CUDA** (NVIDIA CUDA with InfiniBand)
- **MacOSX-x86** (Mac OS X for Intel processors, fails on 10.7 "Lion")
- **MacOSX-x86_64** (Mac OS X for 64-bit Intel processors)
- **MacOSX-PPC** (Mac OS X for PowerPC)
- **Solaris-x86_64**
- **Win32** (Windows XP, etc.)
- **Win64-MPI** (Windows HPC Server)
- **Source Code**

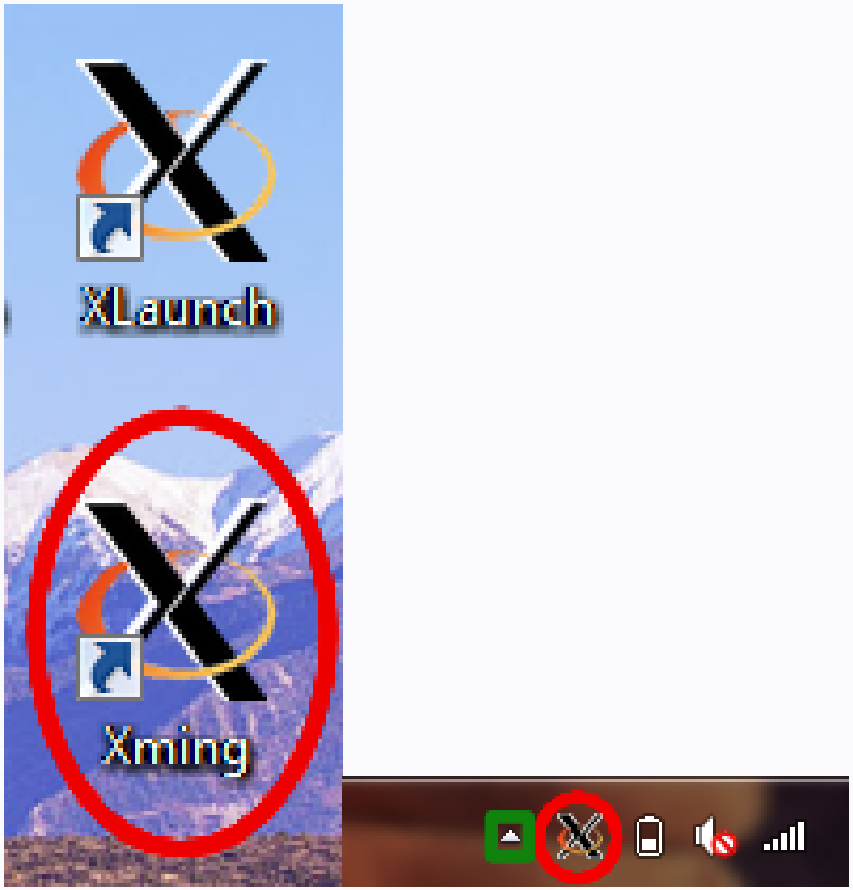
- The amount of installed memory less the amount that is used by the operating system and other utilities
- Max amount per node
  - Linux clusters: ∼6 GB for Queen Bee, ∼3 GB for others
  - Tezpur:∼3 GB
  - Philip:∼22 GB for regular nodes, ∼ 45 or 92 GB for high memory nodes.
  - Pandora:∼125 GB
  - ★ Rule of Thumb: Leave about 1-2 GB for the OS to run

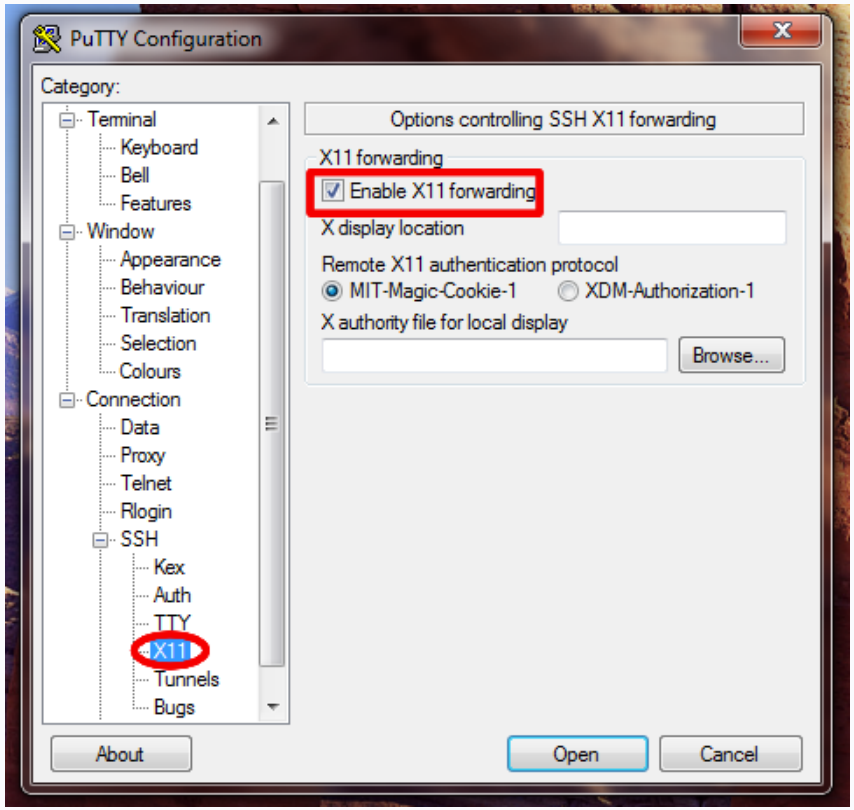
- LONI Host name: `<cluster name>.loni.org`
  - ★ Eric: eric.loni.org
- LSU HPC Host name: `<cluster name>.hpc.lsu.edu`
  - ★ Tezpur: tezpur.hpc.lsu.edu
- Use ssh to connect
  - ★ ∗nix and Mac: `ssh <host name>`
  - ★ Windows: use Putty, Secure Shell Client or Bitvise Tunnelier
- The default Login shell is bash
- Supported shells: bash, tcsh, ksh, csh & sh
- Change the login shell at the profile page
  - ♦ LONI: `https://allocations.loni.org`
  - ♦ LSU HPC: `https://accounts.hpc.lsu.edu`
- Reset your password
  - ♦ LONI: `https://allocations.loni.org/user_reset.php`
  - ♦ LSU HPC: `https://accounts.hpc.lsu.edu/user_reset.php`

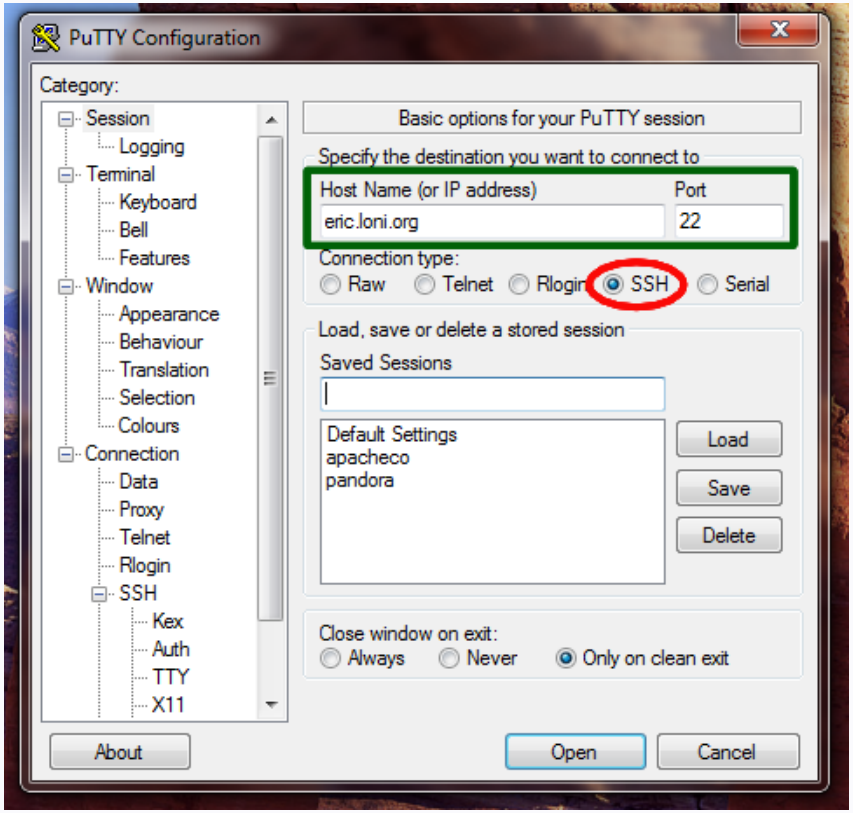

Applications Places System
81 °F Fri Jul 22, 16:13 Alexander Pacheco
eric : etrain00
File Edit View Bookmarks Settings Help
[apacheco@apacheco ~]$ ssh -X etrain00@eric.loni.org
etrain00@eric.loni.org's password:
Last login: Fri Jul 22 15:59:19 2011 from apacheco.cct.lsu.edu
Eric @
Send questions and support is
Eric is a 5 TF, 128 node (512 process
accounts. Eric has its own /home dis
use your /work directory for batch j
limit the number of files per direct
Finally, DO NOT run compute jobs on
terminated.
[etrain00@eric2 ~]$ gview
GaussView 4.1.2 (on eric2)
File Edit View Calculate Results Windows Help
Carbon Tetrahedral
(Default Scheme)
Builder Fragment:
Carbon Tetrahedral
[Hot]
Build Select Placement
eric : etrain00
eric : etrain00
GaussView 4.1.2 (on e...
G1:M1:V1 - New (on e...

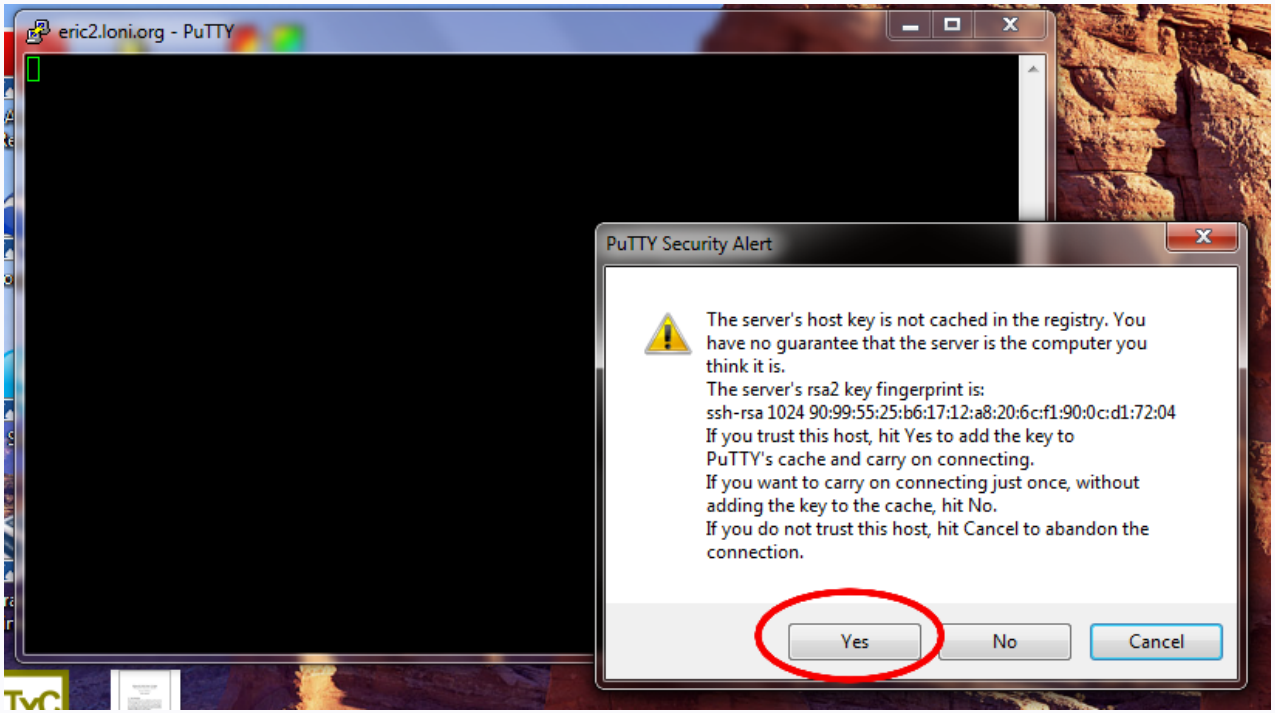

X11 Applications Edit Window Help
(Charged) Fri Jul 22 15:56
etrain00@eric2:~ — ssh — 159×51
etrain00@eric2:~ — ssh
bash
Alex-Pachecos-MacBook-Pro:~ apacheco$ ssh -X etrain00@eric.loni.org
etrain00@eric.loni.org's password:
Warning: No xauth data; using fake authentication data for X11 forwarding.
Last login: Fri Jul 22 07:52:42 2011 from apacheco-1.lsu.edu
Eric @ LSU
Send questions and support issues to: sys-help@loni.org
Eric is a 5 TF, 128 node (512 processor) Linux cluster using LONI LDAP accounts. Eric has its own /home disk. Quotas are set at 5 GB. Please use your /work directory for batch job I/O. It has a 100 GB limit. Please limit the number of files per directory to 10,000.
Finally, DO NOT run compute jobs on the headnodes. Any such jobs will be terminated.
[etrain00@eric2 ~]$ vmd
[1] 27101
[etrain00@eric2 ~]$
VMD Main
File Molecule Graphics Display Mouse Extensions Help
ID T A D F Molecule Atoms Frames Vol
zoom Loop step 1 speed
vmd console
Info) Please include this reference in published work using VMD:
Info) Humphrey, W., Dalke, A. and Schulten, K., `VMD - Visual
Info) Molecular Dynamics', J. Molec. Graphics 1996, 14.1, 33-38.
Info) Multithreading available, 8 CPUs detected.
Info) Free system memory: 7100MB (89%)
Info) OpenGL renderer: NVIDIA GeForce GT 330M OpenGL Engine
Info) Features: STENCIL MSAA(16) MDE MTX NPOT PP PS
Info) GLSL rendering mode is NOT available.
Info) Textures: 2-D (4096x4096), Multitexture (8)
vmd >
VMD 1.8.6 OpenGL Display
VMD
y
x
z

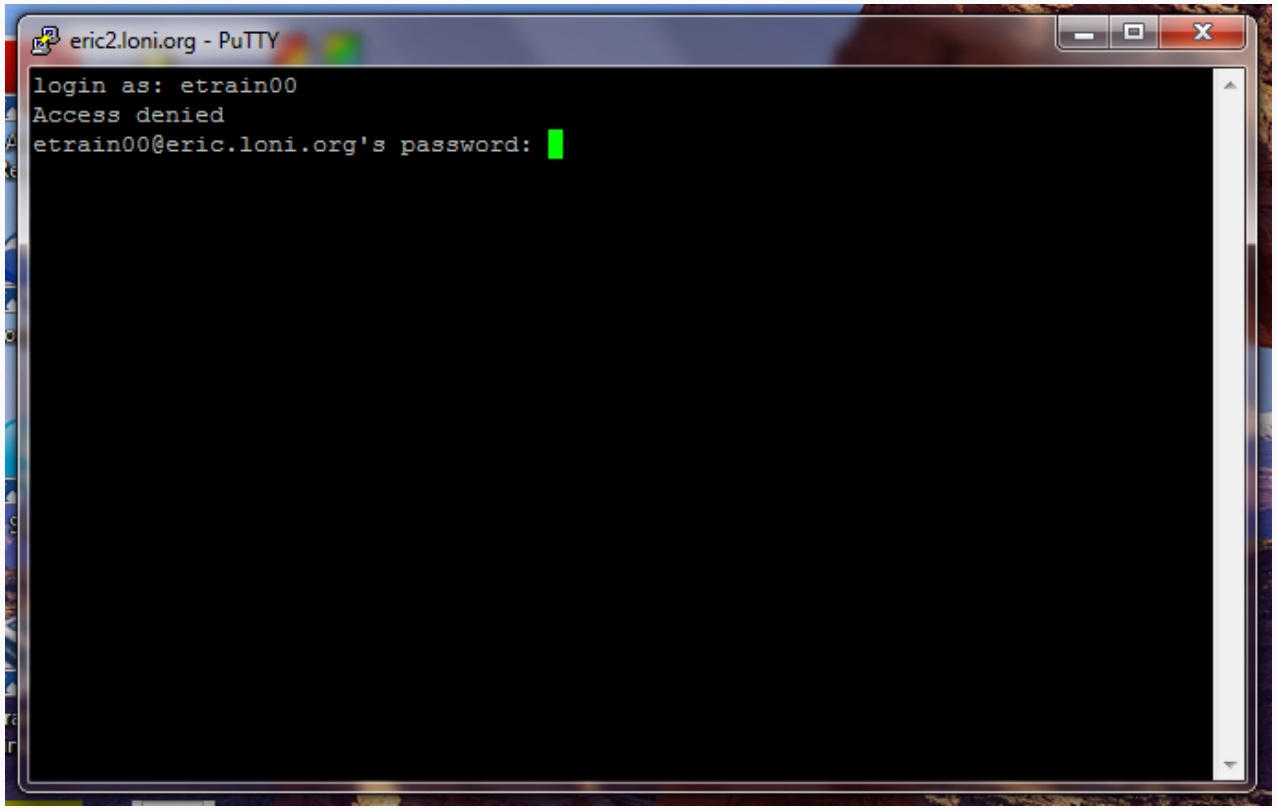
- Download and Install
  1. X-Server: X-ming
     `http://www.straightrunning.com/XmingNotes/`
  2. SSH Client: Putty
     `http://www.chiark.greenend.org.uk/~sgtatham/putty/`
  3. SSH+SFTP/SCP Client: Bitvise Tunnelier
     `http://www.bitvise.com/tunnelier`

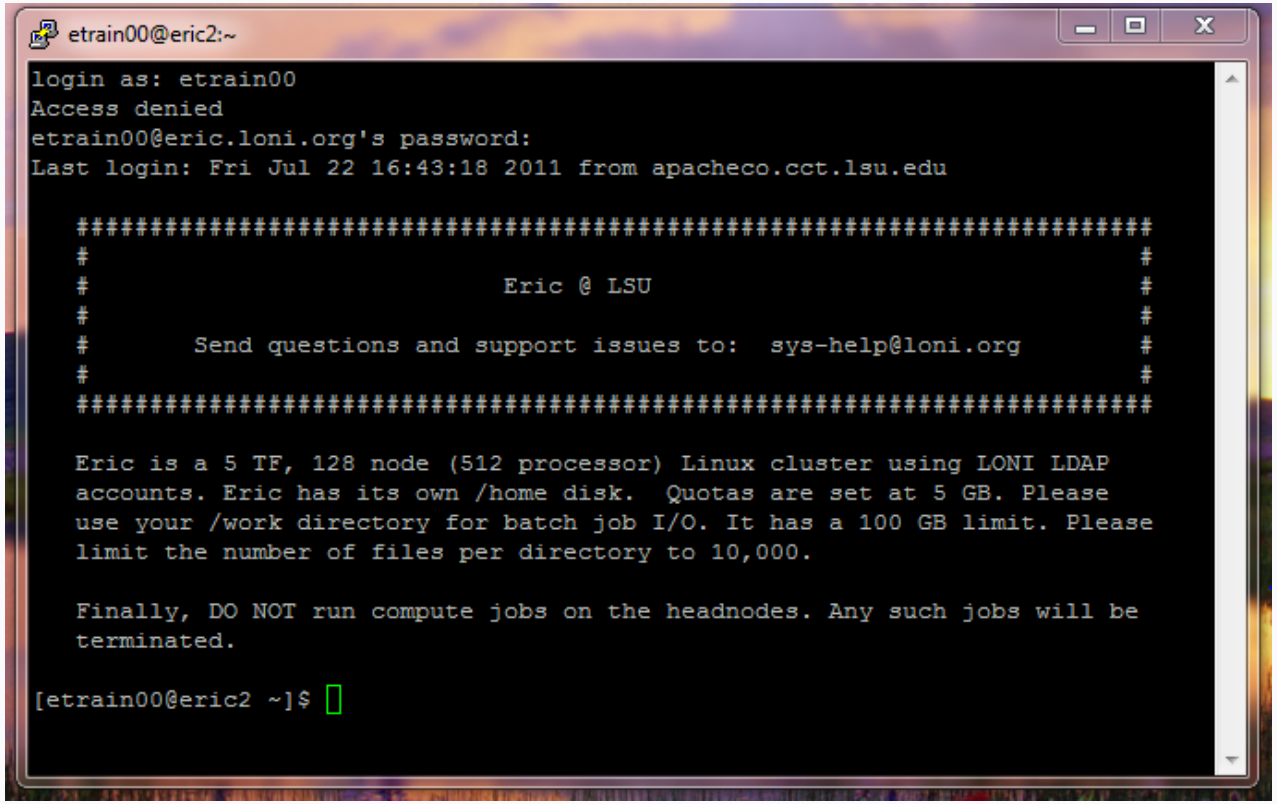

XLaunch
Xming

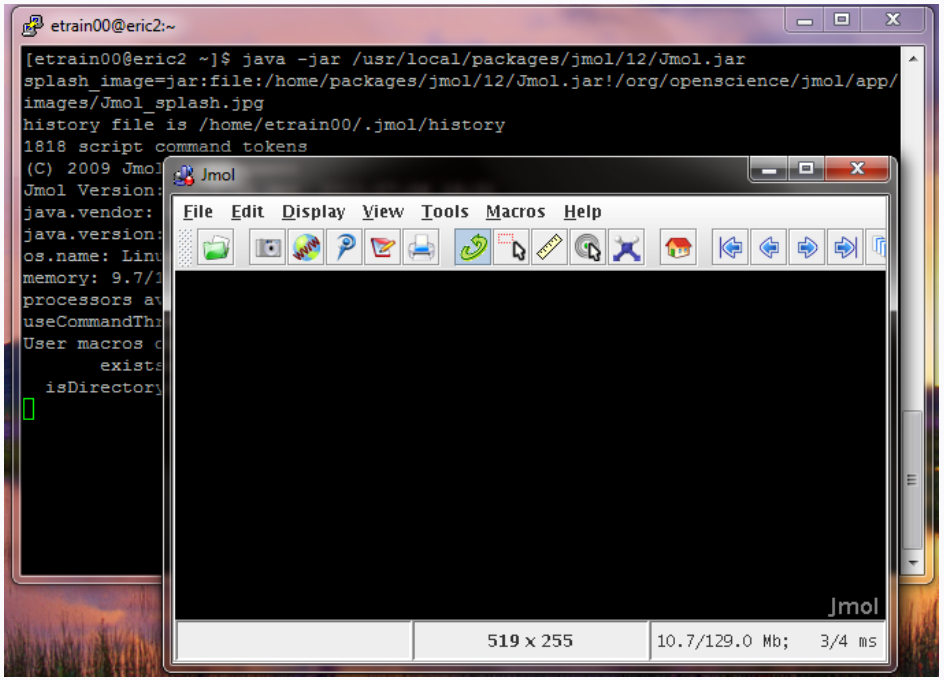

PuTTY Configuration
Category:
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Auth
TTY
X11
Tunnels
Bugs
Options controlling SSH X11 forwarding
X11 forwarding
Enable X11 forwarding
X display location
Remote X11 authentication protocol
MIT-Magic-Cookie-1
XDM-Authorization-1
X authority file for local display
Browse...
About
Open
Cancel

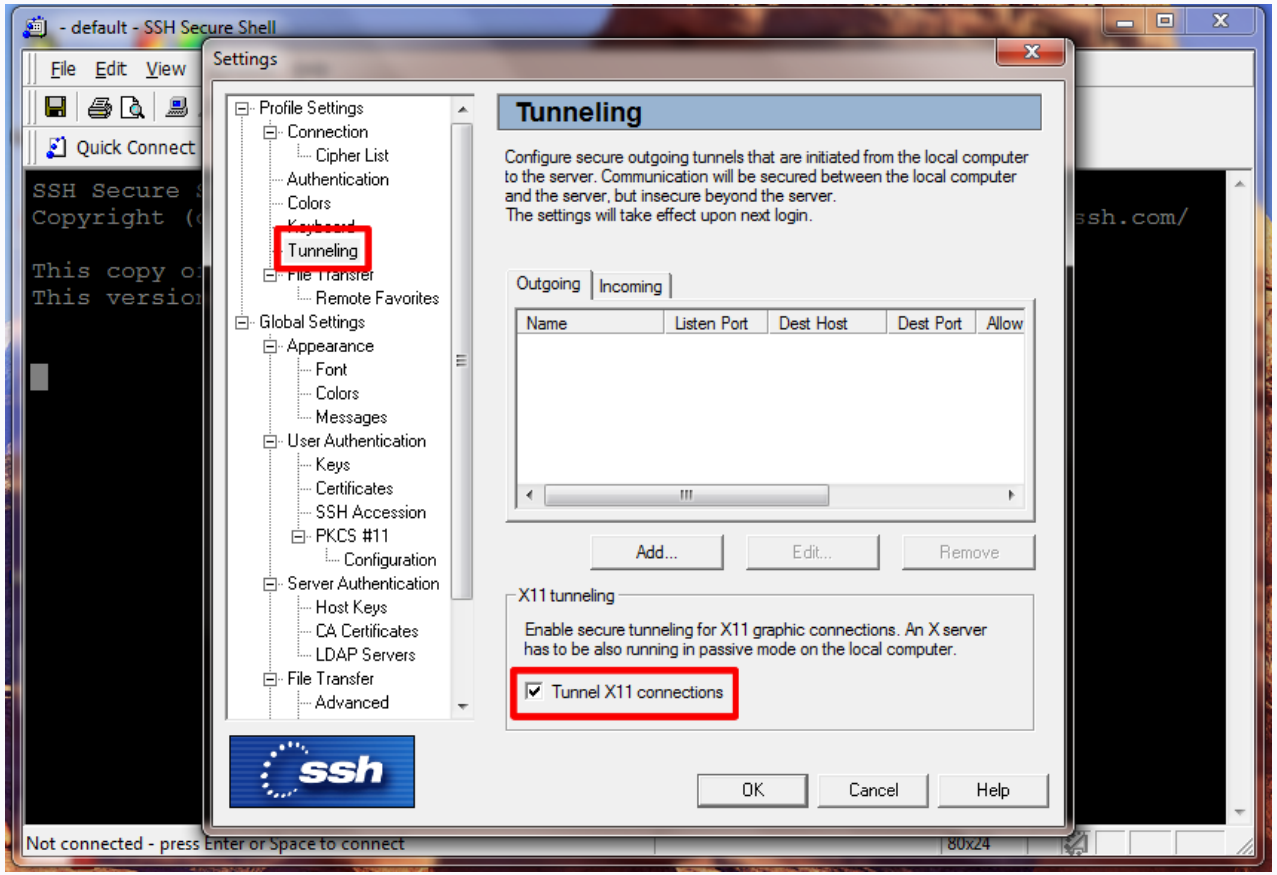

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Auth
TTY
X11
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
eric.loni.org
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
apacheco
pandora
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Open
Cancel

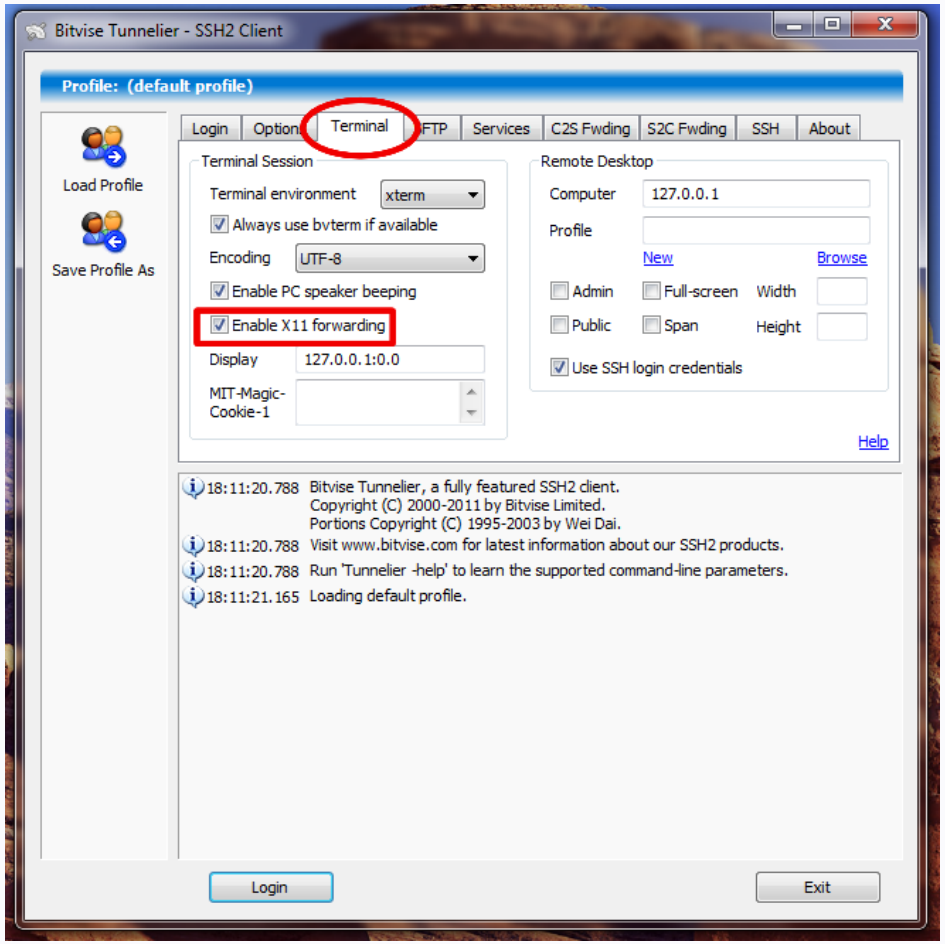

eric2.loni.org - PuTTY
PuTTY Security Alert
The server's host key is not cached in the registry. You have no guarantee that the server is the computer you think it is.
The server's rsa2 key fingerprint is:
ssh-rsa 1024 90:99:55:25:b6:17:12:a8:20:6c:f1:90:0c:d1:72:04
If you trust this host, hit Yes to add the key to PuTTY's cache and carry on connecting.
If you want to carry on connecting just once, without adding the key to the cache, hit No.
If you do not trust this host, hit Cancel to abandon the connection.
Yes
No
Cancel


eric2.loni.org - PuTTY
login as: etrain00
Access denied
etrain00@eric.loni.org's password:


etrain00@eric2:~
login as: etrain00
Access denied
etrain00@eric.loni.org's password:
Last login: Fri Jul 22 16:43:18 2011 from apacheco.cct.lsu.edu
Eric @ LSU
Send questions and support issues to: sys-help@loni.org
Eric is a 5 TF, 128 node (512 processor) Linux cluster using LONI LDAP accounts. Eric has its own /home disk. Quotas are set at 5 GB. Please use your /work directory for batch job I/O. It has a 100 GB limit. Please limit the number of files per directory to 10,000.
Finally, DO NOT run compute jobs on the headnodes. Any such jobs will be terminated.
[etrain00@eric2 ~]$


etrain00@eric2:~
[etrain00@eric2 ~]$ java -jar /usr/local/packages/jmol/12/Jmol.jar
splash_image=jar:file:/home/packages/jmol/12/Jmol.jar!/org/openscience/jmol/app/
images/Jmol_splash.jpg
history file is /home/etrain00/.jmol/history
1818 script command tokens
Jmol
File Edit Display View Tools Macros Help
Jmol
519 x 255
10.7/129.0 Mb; 3/4 ms


- default - SSH Secure Shell
File Edit View
Quick Connect
SSH Secure
Copyright
This copy o
This versio
ssh.com/
Not connected - press Enter or Space to connect
80x24
Settings
Profile Settings
Connection
Cipher List
Authentication
Colors
Keyboard
Tunneling
File Transfer
Remote Favorites
Global Settings
Appearance
Font
Colors
Messages
User Authentication
Keys
Certificates
SSH Accession
PKCS #11
Configuration
Server Authentication
Host Keys
CA Certificates
LDAP Servers
File Transfer
Advanced
Tunneling
Configure secure outgoing tunnels that are initiated from the local computer to the server. Communication will be secured between the local computer and the server, but insecure beyond the server.
The settings will take effect upon next login.
Outgoing
Incoming
Name
Listen Port
Dest Host
Dest Port
Allow
Add...
Edit...
Remove
X11 tunneling
Enable secure tunneling for X11 graphic connections. An X server has to be also running in passive mode on the local computer.
Tunnel X11 connections
ssh
OK
Cancel
Help


Bitvise Tunnelier - SSH2 Client
Profile: (default profile)
Load Profile
Save Profile As
Login
Option
Terminal
FTP
Services
C2S Fwding
S2C Fwding
SSH
About
Terminal Session
Terminal environment
xterm
Always use bvterm if available
Encoding
UTF-8
Enable PC speaker beeping
Enable X11 forwarding
Display
127.0.0.1:0.0
MIT-Magic-Cookie-1
Remote Desktop
Computer
127.0.0.1
Profile
New
Browse
Admin
Full-screen
Width
Public
Span
Height
Use SSH login credentials
Help
18:11:20.788 Bitvise Tunnelier, a fully featured SSH2 client.
Copyright (C) 2000-2011 by Bitvise Limited.
Portions Copyright (C) 1995-2003 by Wei Dai.
18:11:20.788 Visit www.bitvise.com for latest information about our SSH2 products.
18:11:20.788 Run 'Tunnelier -help' to learn the supported command-line parameters.
18:11:21.165 Loading default profile.
Login
Exit

	Distributed File System	Throughput	File life time	Best used for
Home	Yes	Low	Unlimited	Code in development, compiled executable
Work	Yes	High	30 days	Job input/output
Local Scratch	No		Job Duration	Temporary files

- Tips
  - Never write job output to your home directory
  - Do not write temporary files to /tmp, use local scratch or work space
  - Work space is not for long term storage. Files are purged peridocally
  - Use `rmpurge` to delete large amount of files.

Cluster	Home		Work		Scratch
	Access Point	Quota	Access Point	Quota	Access Point
LONI Linux	/home/$USER	5GB	/scratch/$USER	100GB	/var/scratch
HPC Linux	/home/$USER	5GB	/work/$USER	NA	/var/scratch
HPC AIX	/home/$USER	5GB	/work/$USER	50GB	/scratch/local

- No quota is enforced on the work space of QueenBee
- Work directory is created within an hour of first login
- Check current disk usage
- Linux: `showquota`
- AIX: `quota`

Exercise 1

- Log in to any cluster
- Check your disk quota
  1. Linux: `showquota`
  2. AIX: `quota`
- Copy the traininglab directory

```
cp -r /home/apacheco/traininglab .
```

- If you are not familiar with working on a Linux/Unix system
  1. HPC Training Next Week (Sep 19, 2012) Introduction to Linux.
  2. HPC Training Week after next (Sep 26, 2012) Bash Shell Scripting.
  3. Loni Moodle course @ `https://docs.loni.org/moodle`: HPC104 or HPC105.

- Environment variables
  - PATH: where to look for executables
  - LD_LIBRARY_PATH: where to look for shared libraries
  - Other custom environment variables needed by various software
- **SOFTENV** is a software that is used to set up these environment variables on all the clusters
  - More convenient than setting numerous environment variables in .bashrc or .cshrc

- Command `softenv` lists all packages that are managed by **SOFTENV**

```
[apacheco@eric2 ~]$ softenv
SoftEnv version 1.6.2
...
----------------------------------------------------------------------------

These are the macros available:

*   @default


These are the keywords explicitly available:

    +ImageMagick-6.4.6.9-intel-11.1
                                    @types: Applications Visualization @name:
                                     ...
    +NAMD-2.6-intel-11.1-mvapich-1.1
                                    @types: Applications @name: NAMD @version:
                                      ...
    +NAMD-2.7b2-intel-11.1-mvapich-1.1
                                    @types: Applications @name: NAMD @version:
                                      ...
```

- Use `-k` option with `softenv`

```
[apacheco@eric2 ~]$ softenv -k gaussian
SoftEnv version 1.6.2
...
Search Regexp: gaussian
------------------------------------------------------------------------

These are the macros available:

These are the keywords explicitly available:

    +gaussian-03                  @types: Applications Chemistry @name:
                                    Gaussian @version: 03 @build: @internal:
                                    ...
    +gaussian-09                  @types: Applications Chemistry @name:
                                    Gaussian @version: 09 @build: @internal:
                                    ...
    +gaussview-4.1.2              @types: Applications Chemistry @name:
                                    GaussView @version: 4.1.2 @build: - @about:
                                    ...

These are the keywords that are part of the software tree,
however, it is not suggested that you use these:
```

- Setting up environment variables to use a certain package in the current session only.
  - Remove a package: `soft add <key>`
  - Add a package: `soft add <key>`

```
[apacheco@eric2 ~]$ which g09
/usr/local/packages/gaussian09/g09/g09
[apacheco@eric2 ~]$ soft delete +gaussian-09
[apacheco@eric2 ~]$ which g09
/usr/bin/which: no g09 in (/home/apacheco/bin:...
[apacheco@eric2 ~]$ soft add +gaussian-03
[apacheco@eric2 ~]$ which g03
/usr/local/packages/gaussian03/g03/g03
```

- Setting up the environment variables to use a certain software package(s).
  - First add the key to `$HOME/.soft`.
  - Execute `resoft` at the command line.

```
[apacheco@eric2 ~]$ cat .soft
#
# This is the .soft file.
...
+mvapich-1.1-intel-11.1
+intel-fc-11.1
+intel-cc-11.1
+espresso-4.3.1-intel-11.1-mvapich-1.1
+gaussian-09
+lmto-intel-11.1
+nciplot-intel-11.1
+gaussview-4.1.2
+jmol-12
+vmd-1.8.6
+xcrysden-1.5.24-gcc-4.3.2
+tcl-8.5.8-intel-11.1
+gamess-12Jan2009R1-intel-11.1
+nwchem-5.1.1-intel-11.1-mvapich-1.1
+cpmd-3.11.1-intel-11.1-mvapich-1.1
@default
[apacheco@eric2 ~]$ resoft
```

- `soft-dbq` shows which variables are set by a SOFTENV key

```
[apacheco@eric2 ~]$ soft-dbq +espresso-4.3.1-intel-11.1-mvapich-1.1

This is all the information associated with
the key or macro +espresso-4.3.1-intel-11.1-mvapich-1.1.

-----------------------------------------

Name: +espresso-4.3.1-intel-11.1-mvapich-1.1
Description: @types: Applications @name: Quantum Espresso @version: 4.3.1 @build: ...
Flags: none
Groups: none
Exists on: Linux

-----------------------------------------

On the Linux architecture,
the following will be done to the environment:

  The following environment changes will be made:
    ESPRESSO_PSEUDO = /usr/local/packages/espresso/4.3.1/intel-11.1-mvapich-1.1/pseudo
    ESPRESSO_ROOT = /usr/local/packages/espresso/4.3.1/intel-11.1-mvapich-1.1
    ESPRESSO_TMPDIR = /work/${USER}
    PATH = ${PATH}:/usr/local/packages/espresso/4.3.1/intel-11.1-mvapich-1.1/bin

-----------------------------------------
```

Exercise 2: Use Softenv

- Find the key for VISIT (a visualization package).
- Check what variables are set through the key.
- Set up your environment to use VISIT.
- Check if the variables are correctly set by using `which visit`.

Exercise 2: Use Softenv

- Find the key for VISIT (a visualization package).

  ```
  softenv -k visit
  ```

- Check what variables are set through the key.

  ```
  soft-dbq +visit
  ```

- Set up your environment to use VISIT.

  ```
  soft add +visit
  ```

- Check if the variables are correctly set by using `which visit`.

  ```
  /usr/local/packages/visit/bin/visit
  ```

Language	Linux Cluster			AIX Clusters
	Intel	PGI	GNU	XL
Fortran	ifort	pgf77,pgf90	gfortran	xlf,xlf90
C	icc	pgcc	gcc	xlc
C++	icpc	pgCC	g++	xlC

- Usage: <compiler> <options> <your_code>
  - Example: icc -O3 -o myexec mycode.c
- Some compilers options are architecture specific
  - Linux: EM64T, AMD64 or X86_64
  - AIX: power5,power7 or powerpc

Language	Linux Cluster	AIX Clusters
Fortran	mpif77,mpif90	mpxlf,mpxlf90
C	mpicc	mpcc
C++	mpiCC	mpCC

- Usage: <compiler> <options> <your_code>
  - Example: mpif90 -O2 -o myexec mycode.f90
- On Linux clusters
  - Only one compiler for each language
  - There is no intel_mpicc or pg_mpicc
- There are many different versions of MPI compilers on Linux clusters
  - Each of them is built around a specific compiler
  - Intel, PGI or GNU
- It is extremely important to compile and run you code with the same version!!!

- Use the default version if possible
- These MPI compilers are actually wrappers
  - They still use the compilers we've seen on the previous slide
    - Intel, PGI or GNU
  - They take care of everything we need to build MPI codes
    - Head files, libraries etc.
  - What they actually do can be reveal by the `-show` option

```
[apacheco@eric2 ~]$ mpif90 -show
ln -s /usr/local/packages/mvapich/1.1/intel-11.1/include/mpif.h mpif.h
ifort -fPIC -L/usr/local/ofed/lib64 -Wl,-rpath-link -Wl, \
   /usr/local/packages/mvapich/1.1/intel-11.1/lib/shared \
  -L/usr/local/packages/mvapich/1.1/intel-11.1/lib/shared \
  -L/usr/local/packages/mvapich/1.1/intel-11.1/lib \
  -lmpichf90nc -lmpichfarg -lmpich -L/usr/local/ofed/lib64 \
  -Wl,-rpath=/usr/local/ofed/lib64 -libverbs -libumad -lpthread -lpthread -lrt -limf
rm -f mpif.h
```

- Installed under `/usr/local/packages`
- Most of them managed by SOFTENV
  - Numerical and utility libraries
    - FFTW, HDF5, NetCDF, PetSc, Intel MKL
  - Computational Chemistry
    - Amber, CPMD, Gaussian, GAMESS, Gromacs, LAMMPS, NAMD, NWCHEM
  - Visualization
    - GaussView, VisIt, VMD
  - Profiling/debugging tools
    - DDT, Tau, TotalView
  - . . .
- User Requested Packages
  - We install user requested packages either in user's home directory or system wide depending on usage

Exercise 3: Compiling a code

1. Serial Code
   - On Linux cluster, add the soft keys for either Intel (`+intel-fc-11.1`) or GCC (`+gcc-4.3.2`)
   - Compile `hello.f90` with a compiler of your choice
   - Run the executable from the command line
2. Parallel Code
   - On Linux cluster, find the appropriate key for mpi implementation of the above compiler
   - Compile `hello_mpi.f90`
   - Do Not run the parallel code, we'll use a script to submit to a job manager

Exercise 3: Compiling a code

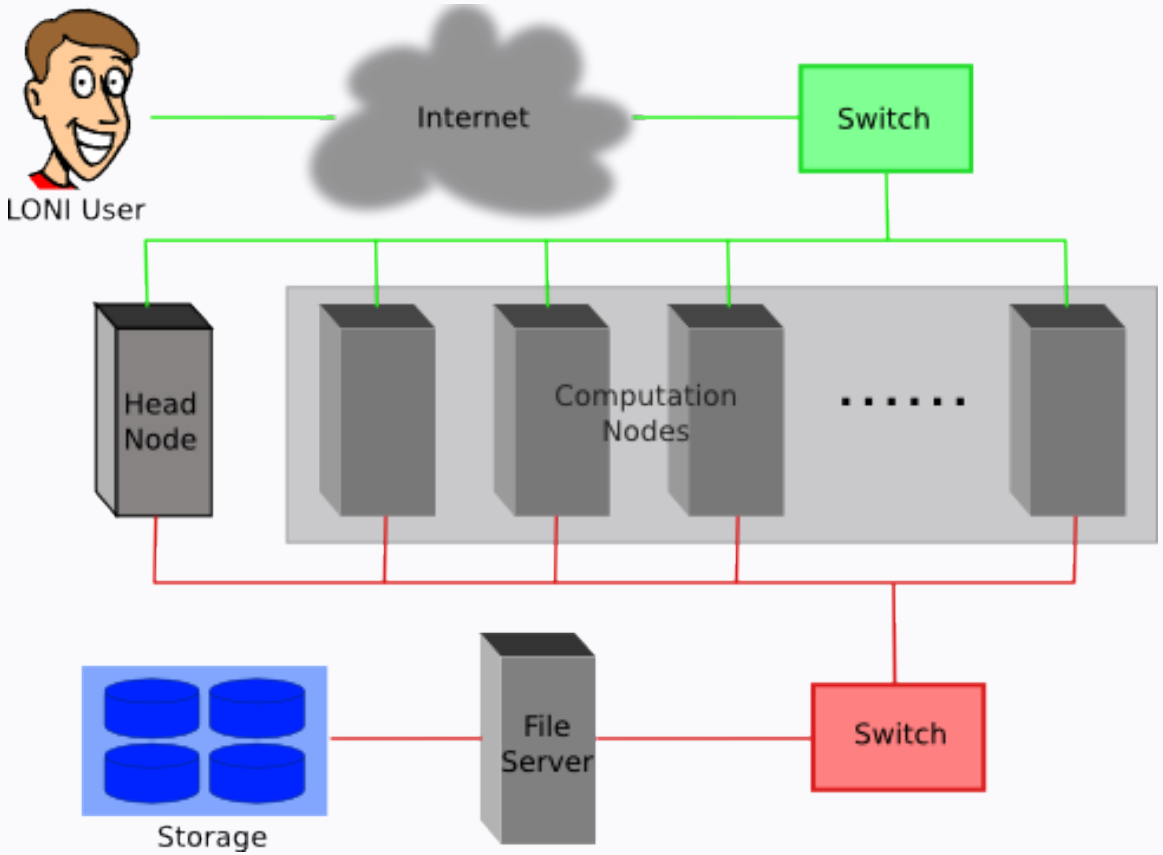
1. Serial Code
   - On Linux cluster, add the soft keys for either Intel (`+intel-fc-11.1`) or GCC (`+gcc-4.3.2`)
   - Compile `hello.f90` with a compiler of your choice

     ```
     ifort -o hello hello.f90
     xlf90 -o hello hello.f90
     ```
   - Run the executable from the command line

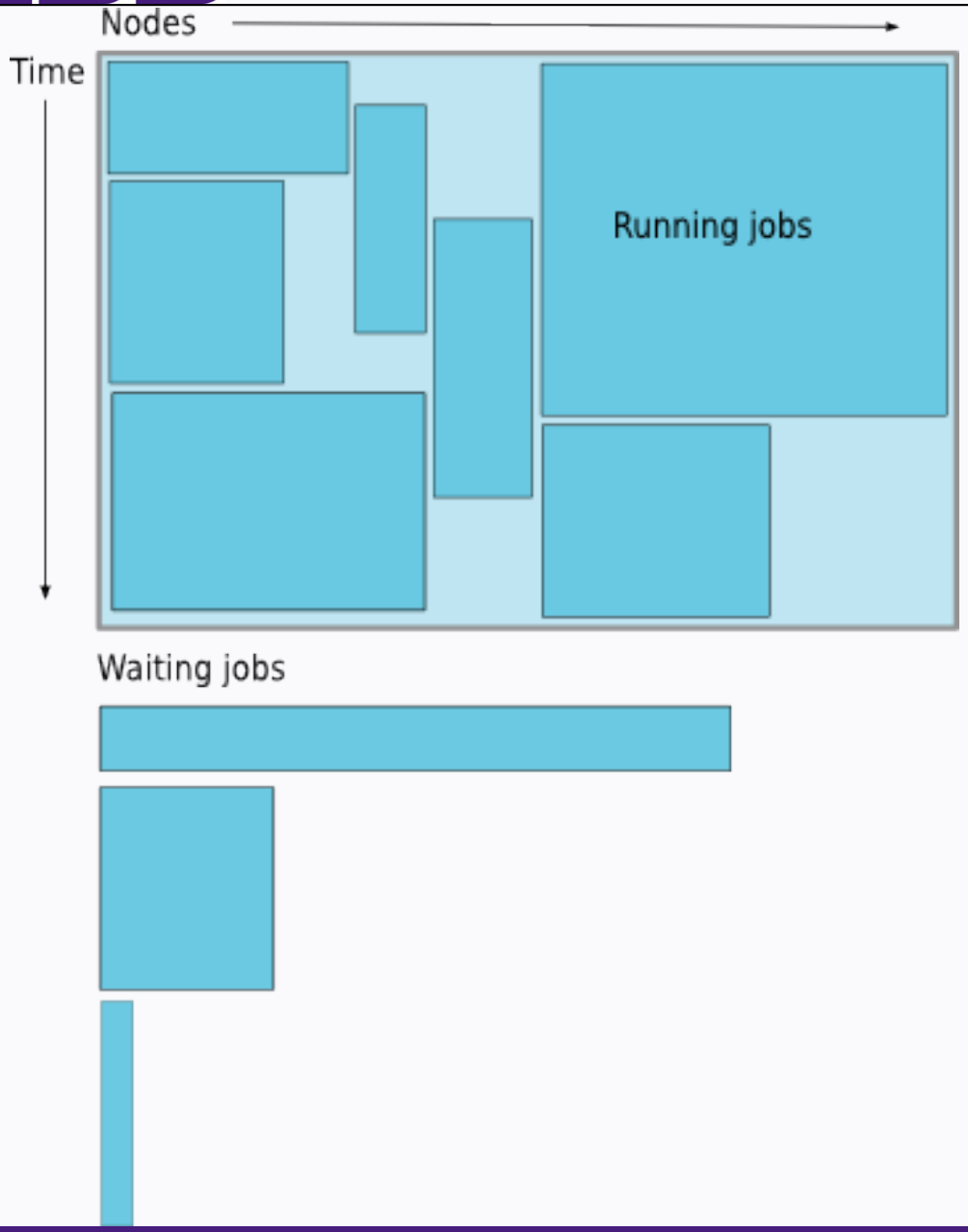
     ```
     ./hello
     ```
2. Parallel Code
   - On Linux cluster, find the appropriate key for mpi implementation of the above compiler
   - Compile `hello_mpi.f90`

     ```
     mpif90 -o hellompi hello_mpi.f90
     ```
   - Do Not run the parallel code, we'll use a script to submit to a job manager

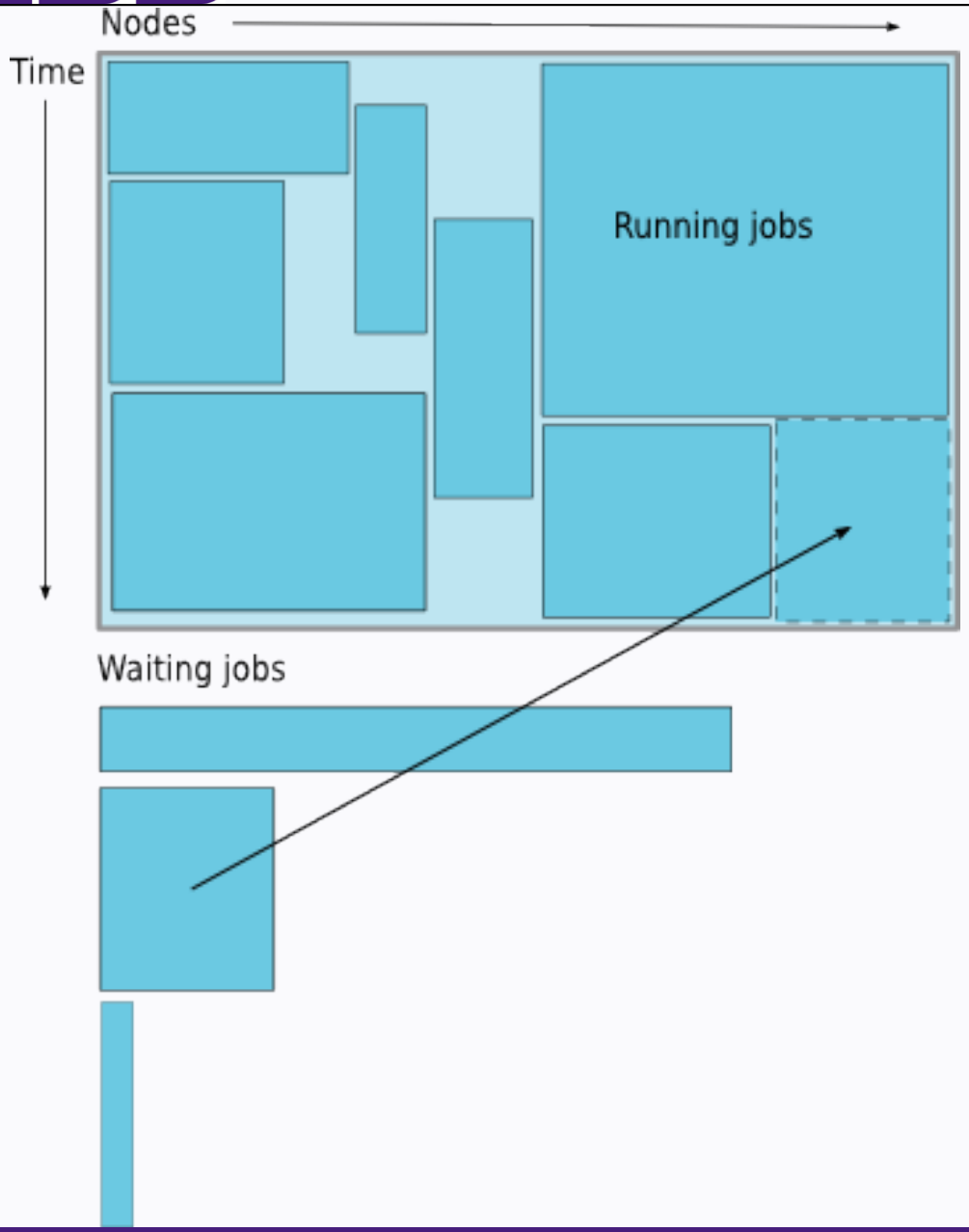
- A cluster is a group of computers (nodes) that works together closely
- Type of nodes
  - Head node
  - Multiple Compute nodes
- Multi User Environment
- Each user may have multiple jobs running simultaneously.



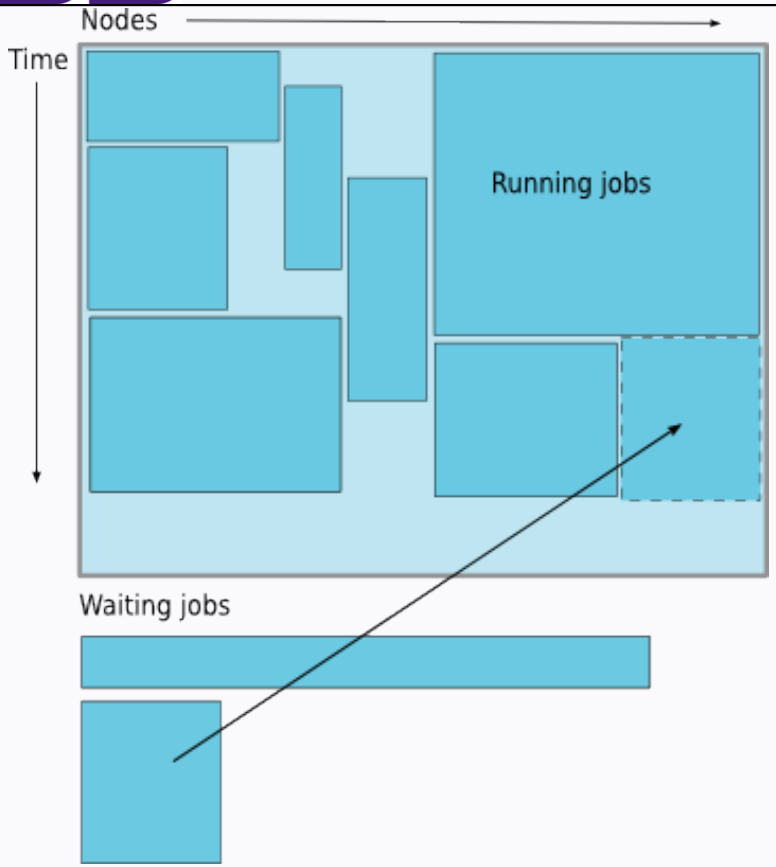
- A software that manages resources (CPU time, memory, etc) and schedules job execution
  - Linux Clusters: Portable Batch System (PBS)
  - AIX Clusters: Loadleveler
- A job can be considered as a user's request to use a certain amount of resources for a certain amount of time
- The batch queuing system determines
  1. The order jobs are executed
  2. On which node(s) jobs are executed

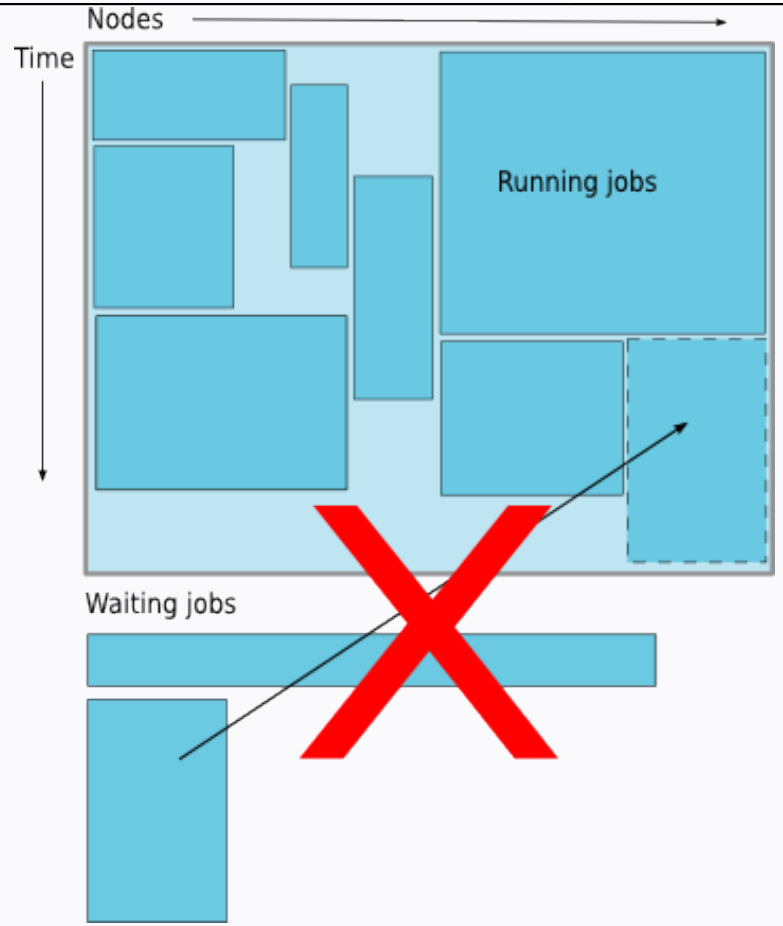


- Map jobs onto the node-time space
  - Assuming CPU time is the only resource
- Need to find a balance between
  - Honoring the order in which jobs are received
  - Maximizing resource utilization



- A strategy to improve utilization
  - Allow a job to jump ahead of others when there are enough idle nodes
  - Must not affect the estimated start time of the job with the highest priority
- Enabled on all LONI and LSU HPC clusters





- Ask for an amount of time that is
  - Long enough for your job to complete
  - As short as possible to increase the chance of backfilling

- There are more than one job queue
- Each job queue differs in
  - Number of available nodes
  - Maximum run time
  - Maximum running jobs per user
- The main purpose is to maximize utilization

QueenBee

Queue	Max Run-time	Total number of nodes	Max running jobs per user	Max nodes per job	Use
workq	3 days	530	8	128	Unpreemptable
checkpt		668		256	preemptable

Other Clusters

Queue	Max Run-time	Total number of nodes	Max running jobs per user	Max nodes per job	Use
single	14 days	16	64	1	Single processor jobs
workq	3 days	96	8	40	Unpreemptable
checkpt		128		64	preemptable

Tezpur

Queue	Max Run-time	Total number of nodes	Max running jobs per user	Max nodes per job	Use
single	3 days	16	64	1	Single processor jobs
workq		180	8	90	Unpreemptable
checkpt		344		180	preemptable

Philip

Queue	Max Run-time	Total number of nodes	Max running jobs per user	Max nodes per job	Use
single	3 days	24	12	1	Single processor jobs
workq		28		5	Unpreemptable
checkpt		28			preemptable
bigmem		5			preemptable
gpu	2 days	2		2	preemptable

Pandora

Queue	Max Run-time	Total number of processors	Max running jobs per user	Max processors per job	Use
interactive	30mins	8	6	8	Interactive Jobs
workq	3 days	224		128	Standard Queue
single	7 days	64		32	Single Node Jobs

- Queue querying
  - Check how busy the cluster is
- Job submission
- Job monitoring
  - Check job status (estimated start time, remaining run time, etc)
- Job manipulation
  - Cancel/Hold jobs

- `qfree`: show number of free,busy and queued nodes
- `qfreeloni`: run `qfree` on all LONI Linux clusters

```
[apacheco@eric2 ~]$ qfree
PBS total nodes: 128,  free: 49,  busy: 79,  down: 0,  use: 61\%
PBS workq nodes: 96,  free: 40,  busy: 28,  queued: 0
PBS checkpt nodes: 104,  free: 40,  busy: 35,  queued: 0
PBS single nodes: 32,  free: 9 *36,  busy: 16,  queued: 366
[apacheco@eric2 ~]$ qfreeloni
-------- qb --------
PBS total nodes: 668,  free: 3,  busy: 647,  down: 18,  use: 96\%
PBS workq nodes: 530,  free: 0,  busy: 278,  queued: 367
PBS checkpt nodes: 668,  free: 1,  busy: 369,  queued: 770
-------- eric --------
PBS total nodes: 128,  free: 49,  busy: 79,  down: 0,  use: 61\%
PBS workq nodes: 96,  free: 40,  busy: 28,  queued: 0
PBS checkpt nodes: 104,  free: 40,  busy: 35,  queued: 0
PBS single nodes: 32,  free: 9 *36,  busy: 16,  queued: 366
-------- louie --------
PBS total nodes: 128,  free: 44,  busy: 83 *2,  down: 1,  use: 64\%
PBS workq nodes: 104,  free: 40,  busy: 0,  queued: 0
PBS checkpt nodes: 128,  free: 44,  busy: 82,  queued: 50
PBS single nodes: 32,  free: 7 *26,  busy: 2,  queued: 0
-------- oliver --------
PBS total nodes: 128,  free: 74,  busy: 52,  down: 2,  use: 40\%
PBS workq nodes: 62,  free: 8,  busy: 11,  queued: 0
...
```

- Command: `llclass`

```
pandora1:~ apacheco$ llclass
Name                 MaxJobCPU     MaxProcCPU  Free   Max Description
                    d+hh:mm:ss     d+hh:mm:ss Slots Slots
-------------- -------------- -------------- ----- ----- --------------------
interactive          unlimited      unlimited     8     8 Queue for interactive jobs; maximum runt
workq                unlimited      unlimited    28   224 Standard queue for job submissions; maxi
cheme                unlimited      unlimited    24    96 Queue for Chemical Engineering; maximum
single               unlimited      unlimited    24    64 Queue for single-node job submissions; m
--------------------------------------------------------------------------------
"Free Slots" values of the classes "workq", "cheme" are constrained by the MAX_STARTERS limit(s).
```

Interactive Jobs

- Set up an interactive environment on compute nodes for users
  - Advantage: can run programs interactively
  - Disadvantage: must be present when job starts
- Purpose: testing and debugging code. **Do not run jobs on head node!!!**

```
qsub -I -V -l walltime=<hh:mm:ss>,nodes=<# of
nodes>:ppn=cpu -A <your allocation> -q <queue name>
```

- On QueenBee, cpu=8
- Other LONI Clusters: cpu=4 (parallel jobs)
- To enable X-forwarding: add `-X`

Batch Jobs

- Executed using a batch script without user intervention
  - Advantage: system takes care of running the job
  - Disadvantage: can change sequence of commands after submission
- Useful for Production runs

```
qsub <job script>
```

```
llsubmit <job script>
```

```
#!/bin/bash
#PBS -l nodes=4:ppn=4
#PBS -l walltime=24:00:00
#PBS -N myjob
#PBS -o <file name>
#PBS -e <file name>
#PBS -q checkpt
#PBS -A <loni_allocation>
#PBS -m e
#PBS -M <email address>

<shell commands>
mpirun  -machinefile $PBS_NODEFILE \
 -np 16 <path_to_executable> <options>
<shell commands>
```

```
Shell being used
# of nodes & processors
Maximum walltime
Job name
standard output
standard error
Queue name
Allocation name
Send mail when job ends
to this address

shell commands
run parallel job

shell commands
```

```
#!/bin/bash
#PBS -l nodes=1:ppn=1
#PBS -l walltime=24:00:00
#PBS -N myjob
#PBS -o <file name>
#PBS -e <file name>
#PBS -q single
#PBS -A <loni_allocation>
#PBS -m e
#PBS -M <email address>

<shell commands>
<path_to_executable> <options>
<shell commands>
```

```
Shell being used
# of nodes & processors
Maximum walltime
Job name
standard output
standard error
Use single queue
Allocation name
Send mail when job ends
to this address

shell commands
run parallel job
shell commands
```

Loadleveler Job Script I


```
#!/bin/sh
#@ job_type = parallel
#@ output = $(jobid).out
#@ error = $(jobid).err
#@ notification = error
#@ notify_user = youremail@domain
#@ class = checkpt
#@ wall_clock_limit = 24:00:00
#@ node_usage = shared
#@ node = 2
#@ total_tasks = 32
#@ resources = ConsumableMem(512 mb) ConsumableCPUS(1)
#@ network.MPI_LAPI = sn_all,shared,US.HIGH
#@ requirements = (Arch == "POWER7")
#@ environment = COPY_ALL
#@ queue

<shell commands>
poe <path_to_executable> <options>
<shell commands>
```

```
Shell being used
Job Type
standard output
standard error
notify on error
to mail address
job queue
max walltime
node usage
# of nodes
total processors
resources requested per processor
network directive
job requirements
environment


shell commands
run parallel job
shell commands
```

```
#!/bin/sh
#@ job_type = serial
#@ output = $(jobid).out
#@ error = $(jobid).err
#@ notification = error
#@ notify_user = youremail@domain
#@ class = checkpt
#@ wall_clock_limit = 24:00:00
#@ node_usage = shared
#@ node = 1
#@ total_tasks = 1
#@ resources = ConsumableMem(512 mb) ConsumableCPUS(1)
#@ requirements = (Arch == "POWER7")
#@ environment = COPY_ALL
#@ queue

<shell commands>
<path_to_executable> <options>
<shell commands>
```

```
Shell being used
Job Type
standard output
standard error
notify on error
to mail address
job queue
max walltime
node usage
# of nodes
total processors
resources requested per processor
job requirements
environment


shell commands
run parallel job
shell commands
```

- On Pandora:
  - `#@ resources = ConsumableMem(512 mb)` `ConsumableCPUS(1)` is required
  - `#@ requirements = (Arch == "POWER7")`
  - `#@ network.MPI_LAPI = sn_all,shared,US,HIGH`

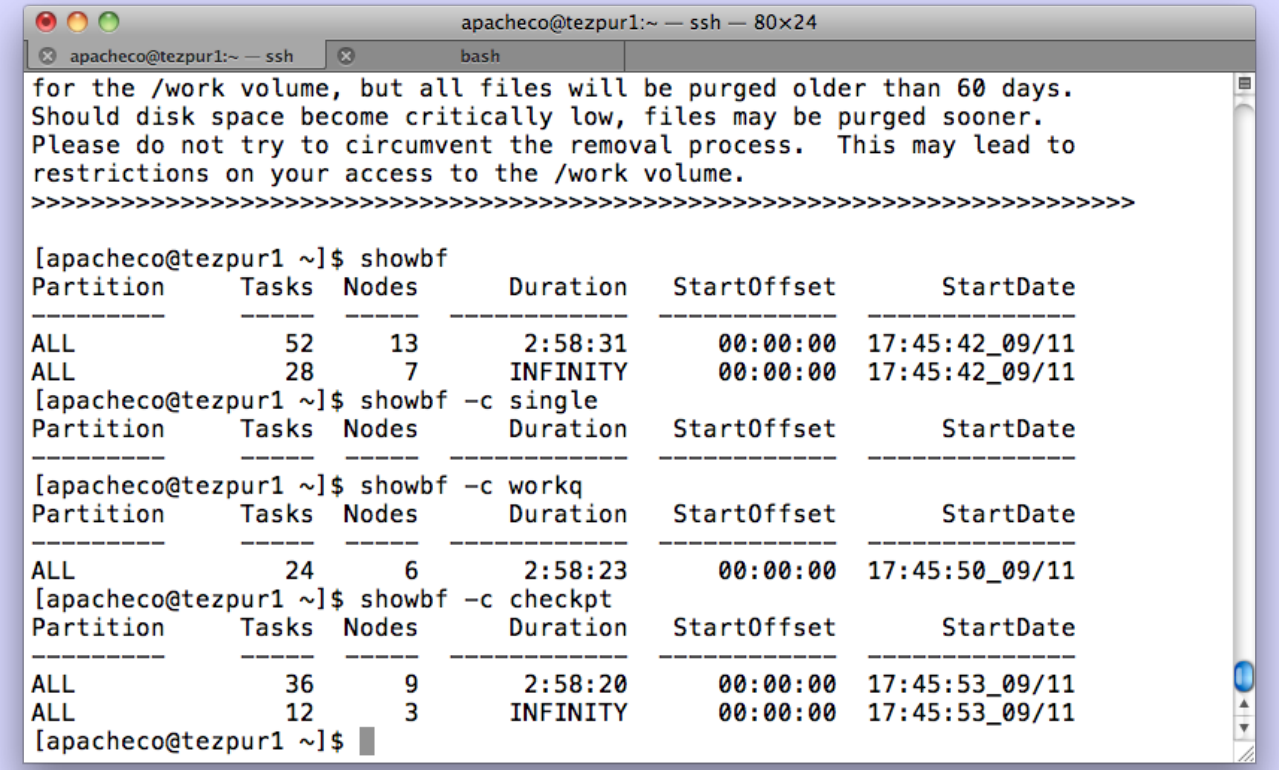
Exercise 4: Job Submission

- Write a job submission script to execute the `hello_mpi` program.
- Submit the script to the job manager.

Linux Clusters

- `showstart <job id>`
  - Check estimated time when job can start
- When can the estimated time change
  - Higher priority job gets submitted
  - Running jobs terminate earlier than time requested
  - System has trouble starting your job
- `qstat <options> <job id>`
  - Show information on job status
  - All jobs displayed if `<job id>` is omitted
  - `qstat -u <username>`: Show jobs belonging to `<username>`
  - `qstat -a <job id>`: Display in an alternative format
- `qshow <job id>`
  - Show information of running job `<job id>`: node running on and CPU load

- `showbf -c <queue>`



AIX Clusters

- `llq <options> <job id>`
  - All jobs are displayed if `<job id>` is omitted
  - Display detailed information: `llq -l <job id>`
  - Check estimated start time: `llq -s <job id>`
  - Show jobs from a specific user: `llq -u <username>`

```
pandora1:~ apacheco$ llq
Id                       Owner      Submitted   ST PRI Class        Running On
------------------------ ---------- ----------- -- --- ------------ -----------
pandora1.12700.0         ghoshbd     9/9  13:59 R  50  single       pandora002
pandora1.12701.0         ghoshbd     9/9  14:02 R  50  workq        pandora007
pandora1.12702.0         ghoshbd     9/9  14:04 R  50  single       pandora001
pandora1.12713.0         abhi300     9/10 17:12 R  50  single       pandora001
pandora1.12714.0         ghoshbd     9/11 01:24 R  50  workq        pandora004
pandora1.12716.0         hypoxia     9/11 09:14 R  50  workq        pandora008
pandora1.12717.0         abhi300     9/11 09:57 R  50  single       pandora002

7 job step(s) in queue, 0 waiting, 0 pending, 7 running, 0 held, 0 preempted
```

AIX Clusters

- `showllstatus.py`: Show job status and node running on

```
pandora1:~ apacheco$ showllstatus.py
Node          Status   Load       Arch Node          Status   Load       Arch
pandora001    Run      20.20    POWER7 pandora006    Busy     32.09    POWER7
pandora002    Run      20.39    POWER7 pandora007    Run      16.01    POWER7
pandora003    Busy     32.04    POWER7 pandora008    Busy     32.10    POWER7
pandora004    Busy     32.07    POWER7 pandora1      Idle     0.00     POWER7
pandora005    Busy     32.02    POWER7

Step ID             Owner        Status   Class    Hosts  Queue Date   Disp. Date
pandora1.12717.0    abhi300         R     single     1    09/11 09:57  09/11 09:57
pandora1.12716.0    hypoxia         R     workq      4    09/11 09:14  09/11 09:14
pandora1.12714.0    ghoshbd         R     workq      1    09/11 01:24  09/11 01:24
pandora1.12713.0    abhi300         R     single     1    09/10 17:12  09/10 17:12
pandora1.12702.0    ghoshbd         R     single     1    09/09 14:04  09/09 14:04
pandora1.12701.0    ghoshbd         R     workq      1    09/09 14:02  09/09 14:02
pandora1.12700.0    ghoshbd         R     single     1    09/09 13:59  09/09 13:59
```

Linux Clusters

- `qdel <job id>` — Cancel a running or queued job
- `qhold <job id>` — Put a queued job on hold
- `qrls <job id>` — Resume a held job

AIX Clusters

- `llcancel <job id>` — Cancel a running or queued job
- `llhold <job id>` — Put a queued job on hold
- `llhold -r <job id>` — Resume a held job

- User's Guide
  - HPC: `http://www.hpc.lsu.edu/help`
  - LONI: `https://docs.loni.org/wiki/Main_Page`
- Contact us
  - Email ticket system: sys-help@loni.org
  - Telephone Help Desk: 225-578-0900
  - Walk-in consulting session at Middleton Library
    - ★ Tuesdays and Thursdays only
  - Instant Messenger (AIM, Yahoo Messenger, Google Talk)
    - ★ Add "lsuhpchelp"