Intro to NCAR HPC Resources

2022 CESM Tutorial

Rory Kelly

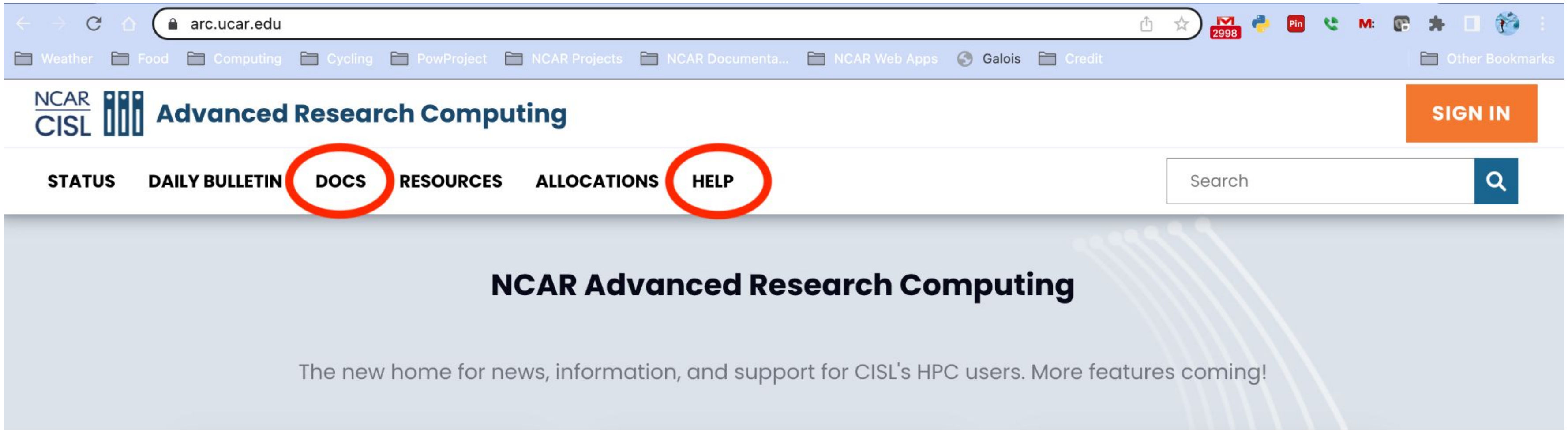
Consulting Services Group, CISL

August 08, 2022

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Getting Help

https://arc.ucar.edu/



NCAR Advanced Research Computing

The new home for news, information, and support for CISL's HPC users. More features coming!

- Searchable Documentation
- Support Tickets with HPC Consultants

Topics to Cover

- Available systems and their uses
- Signing in and managing data
- Accessing and building software
- Managing jobs using Batch schedulers
- Customizing your user environment

Cheyenne - Model runs, Large Parallel Jobs

Cheyenne

SGI ICE XA Supercomputer

Entered production January 2017

4032 Compute nodes (145,152 total cores)

- Dual socket, 18 cores per socket

  2.3 GHz Intel Xeon (Broadwell) processors

  313 TB total system memory, DDR4-2400

  - 64 GB/node, single-rank DIMM, 3168 nodes
  - 128 GB/node, dual-rank DIMM, 864 nodes

  Mellanox EDR InfiniBand, Partial 9D Enhanced Hypercube Topology

6 login nodes

- Dual socket, 18 cores per socket, 256 GB memory/node

Casper - Data Analysis, Visualization, Machine Learning, GPU Computing, HTC

Casper

Heterogeneous cluster of specialized nodes targeting data analysis, visualization, and GPU computing.

- 22 Supermicro SuperWorkstation nodes are used for data analysis and visualization jobs. Each node has 36 cores and up to 384 GB memory.
  - 9 of these nodes also feature an NVIDIA Quadro GP100 GPU.
- 10 nodes feature large-memory, dense GPU configurations to support explorations in machine learning (ML) and GPU computing
  - 4 of these nodes feature 4 NVIDIA Tesla V100 GPUs
  - 6 of these nodes feature 8 NVIDIA Tesla V100 GPUs
- 64 high-throughput computing (HTC) nodes for small computing tasks using 1 or 2 CPUs.
  - 62 HTC nodes have 384 GB of available memory
  - 2 HTC nodes have 1.5 TB of available memory
- 4 nodes are reserved for Research Data Archive workflows.

Next HPC System - Derecho

Derecho
HPE Cray EX Supercomputer
Delivery in Q4 2022

2488 CPU Compute nodes (318,464 total cores)

- Dual socket, 64-core AMD Milan processors
- 256 GB DDR4 memory

82 GPU Compute nodes (5248 CPU cores + 328 GPUs)

- Single socket, 64-core AMD Milan processor
- 4 A100 GPUs, 40GB HBM2 memory per GPU
- 512 GB DDR4 memory

HPE Slingshot 11 Interconnect

- Dragonfly topology
- 200 Gb/sec per port per direction
- 1.7 - 2.6 μs latency
- Adaptive routing and congestion control
- 1 injection port on CPU nodes, 4 ports on GPU nodes

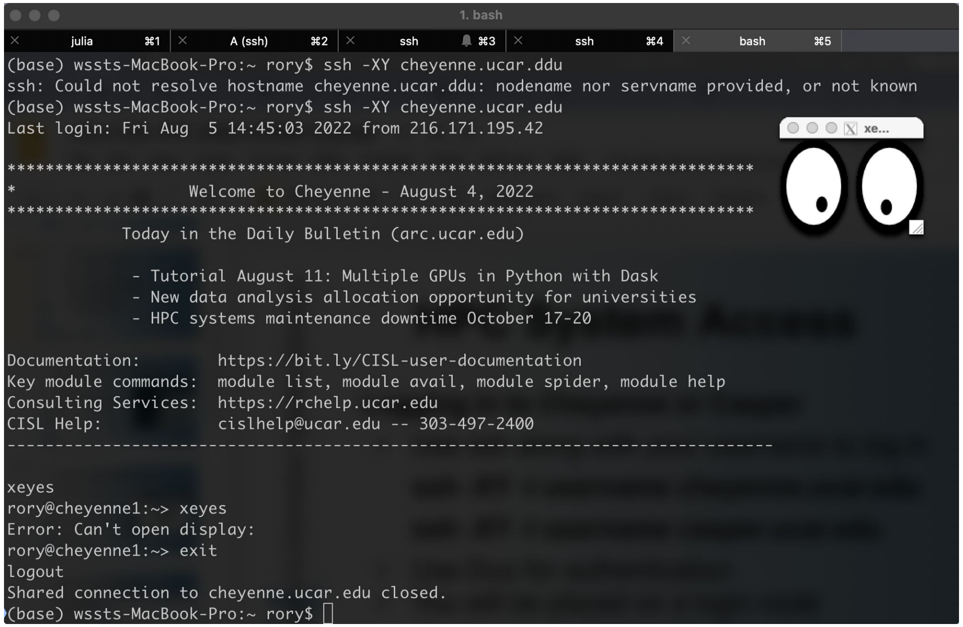
HPC System Access - Logging in

Use ssh along with your username to log in

ssh -XY -l username cheyenne.ucar.edu

ssh -XY -l username casper.ucar.edu

- Use Duo for authentication
- Cheyenne - 6 login nodes
- Casper - 2 login nodes



```
(base) wssts-MacBook-Pro:~ rory$ ssh -XY cheyenne.ucar.ddu
ssh: Could not resolve hostname cheyenne.ucar.ddu: nodename nor servname provided, or not known
(base) wssts-MacBook-Pro:~ rory$ ssh -XY cheyenne.ucar.edu
Last login: Fri Aug  5 14:45:03 2022 from 216.171.195.42

******************************************************************************
*                  Welcome to Cheyenne - August 4, 2022
******************************************************************************
            Today in the Daily Bulletin (arc.ucar.edu)

             - Tutorial August 11: Multiple GPUs in Python with Dask
             - New data analysis allocation opportunity for universities
             - HPC systems maintenance downtime October 17-20

Documentation:        https://bit.ly/CISL-user-documentation
Key module commands:  module list, module avail, module spider, module help
Consulting Services:  https://rchelp.ucar.edu
CISL Help:            cislhelp@ucar.edu -- 303-497-2400
--------------------------------------------------------------------------------

xeyes
rory@cheyenne1:~> xeyes
Error: Can't open display:
rory@cheyenne1:~> exit
logout
Shared connection to cheyenne.ucar.edu closed.
(base) wssts-MacBook-Pro:~ rory$ 
```

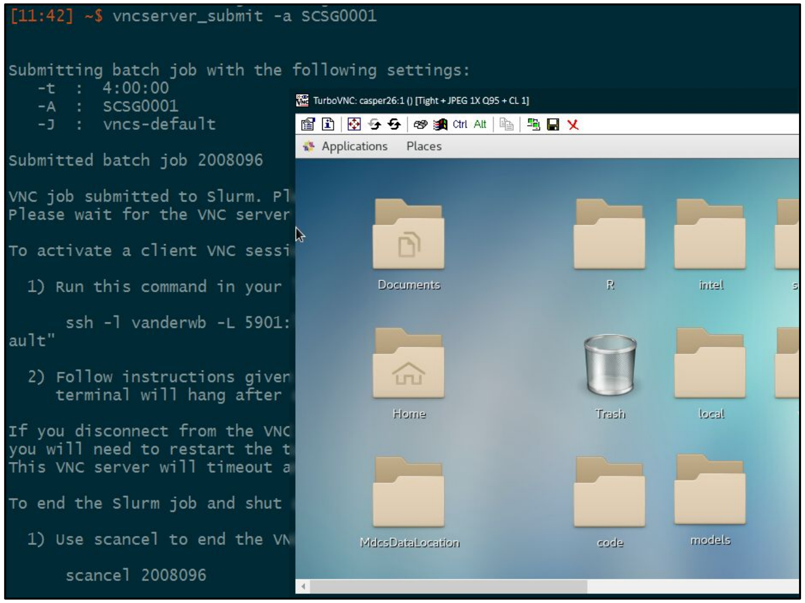
Run GUI Programs with TigerVNC

VNC can be used to run a remote GNOME/KDE desktop

Need to install a VNC client first - We recommend TigerVNC, but other VNC clients such as TurboVNC will also work

Usage:

```
vncserver_submit -a <project>
```



```
[11:42] ~$ vncserver_submit -a SCSG0001

Submitting batch job with the following settings:
   -t  :  4:00:00
   -A  :  SCSG0001
   -J  :  vncs-default

Submitted batch job 2008096

VNC job submitted to Slurm. Pl
Please wait for the VNC server

To activate a client VNC sessi

  1) Run this command in your

      ssh -l vanderwb -L 5901:
ault"

  2) Follow instructions given
     terminal will hang after

If you disconnect from the VNC
you will need to restart the t
This VNC server will timeout a

To end the Slurm job and shut

  1) Use scancel to end the VN

      scancel 2008096
```

Run GUI Programs with FastX

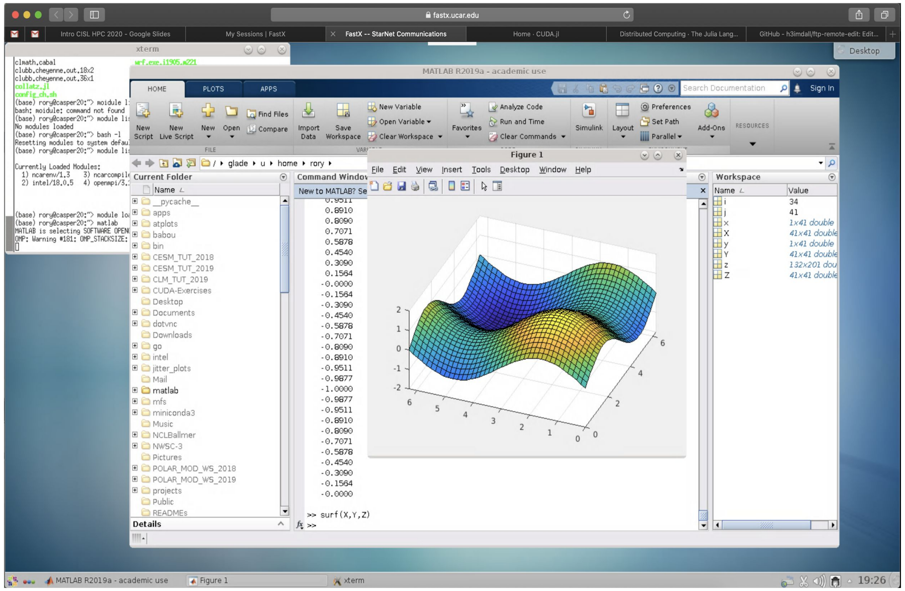
FastX can also be used to run a remote desktop or terminal session

Can be run in a browser without a client

Connect to the NCAR VPN, then go to

https://fastx.ucar.edu:3300

Can also be setup using an SSH tunnel or a desktop client



```
clmath.cabal
clubb.cheyenne.out.18x2
clubb.cheyenne.out.36x1
collatz.jl
config_ch.sh
(base) rory@casper20:~> moidule li
bash: moidule: command not found
(base) rory@casper20:~> module lis
No modules loaded
(base) rory@casper20:~> bash -l
Resetting modules to system defaul
(base) rory@casper20:~> module lis

Currently Loaded Modules:
  1) ncarenv/1.3    3) ncarcompile
  2) intel/18.0.5   4) openmpi/3.1



(base) rory@casper20:~> module loa
(base) rory@casper20:~> matlab
MATLAB is selecting SOFTWARE OPENG
OMP: Warning #181: OMP_STACKSIZE:
```

```
    0.9511
    0.8910
    0.8090
    0.7071
    0.5878
    0.4540
    0.3090
    0.1564
   -0.0000
   -0.1564
   -0.3090
   -0.4540
   -0.5878
   -0.7071
   -0.8090
   -0.8910
   -0.9511
   -0.9877
   -1.0000
   -0.9877
   -0.9511
   -0.8910
   -0.8090
   -0.7071
   -0.5878
   -0.4540
   -0.3090
   -0.1564
   -0.0000

>> surf(X,Y,Z)
>>
```

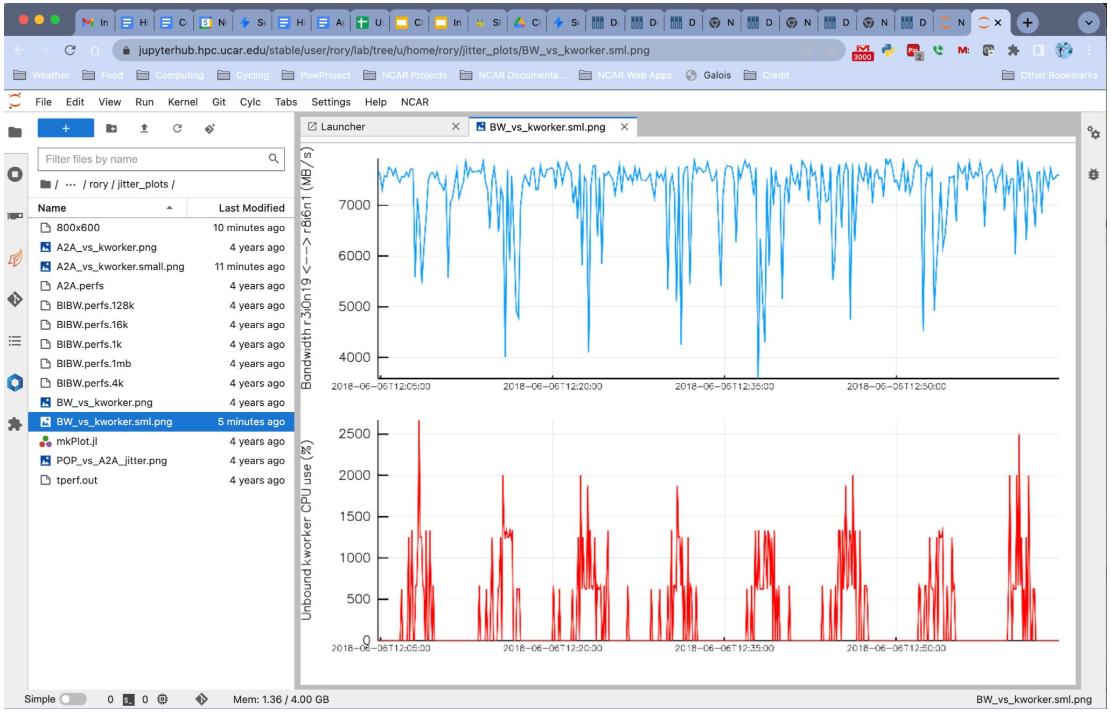
Name	Value
i	34
j	41
x	1x41 double
X	41x41 double
y	1x41 double
Y	41x41 double
z	132x201 dou
Z	41x41 double

View images with jupyterhub

https://jupyterhub.hpc.ucar.edu

Jupyterhub is used to create sophisticated interactive computational notebooks for analysis, education, etc.

It can also be used for unsophisticated things like viewing images



Name	Last Modified
800x600	10 minutes ago
A2A_vs_kworker.png	4 years ago
A2A_vs_kworker.small.png	11 minutes ago
A2A.perfs	4 years ago
BIBW.perfs.128k	4 years ago
BIBW.perfs.16k	4 years ago
BIBW.perfs.1k	4 years ago
BIBW.perfs.1mb	4 years ago
BIBW.perfs.4k	4 years ago
BW_vs_kworker.png	4 years ago
BW_vs_kworker.sml.png	5 minutes ago
mkPlot.jl	4 years ago
POP_vs_A2A_jitter.png	4 years ago
tperf.out	4 years ago

Be mindful when using shared login nodes

- Your activities coexists with those of other users
- CPUs and memory are shared on the login nodes
- Limit your usage to:
  - Reading and editing text/code
  - Compiling small programs
  - Performing data transfers
  - Interacting with the job scheduler
- Programs that use excessive resources on the login nodes will be terminated

Data storage - GLADE

File spaces optimized for parallel IO, accessible from all HPC systems

File space	Quota	Backup	Uses
Home **/glade/u/home/$USER**	50 GB	Yes	Settings, code, scripts
Work **/glade/work/$USER**	1 TB	No	Compiled codes, models
Scratch **/glade/scratch/$USER**	10 TB	**Purged!**	Run directories, temp output
Project **/glade/p/entity/project_code**	N/A	No	Project space allocations

*Keep track of usage with “**gladequota**”*

Data storage - Campaign Storage

- Resource for storing data on project allocation time scales (3-5 years)
- Data expected to be migrated after 5 years.
- Multiple access methods
  - Globus (NCAR Campaign Storage)
  - Casper nodes (/glade/campaign/)
  - Data access nodes (/glade/campaign/)
- Allocated to and managed by NCAR labs and can be requested by University users

Data storage - Collections

- Curated data collections available on Cheyenne and Casper to facilitate easy access to research data sets
- RDA
  - Research Data Archive
  - /glade/collections/rda/
  - https://www2.cisl.ucar.edu/data-portals/research-data-archive
- CMIP6
  - Coupled Model Intercomparison Project
  - /glade/collections/cmip/CMIP6/
  - https://www2.cisl.ucar.edu/resources/cmip-analysis-platform

Moving data to and from GLADE

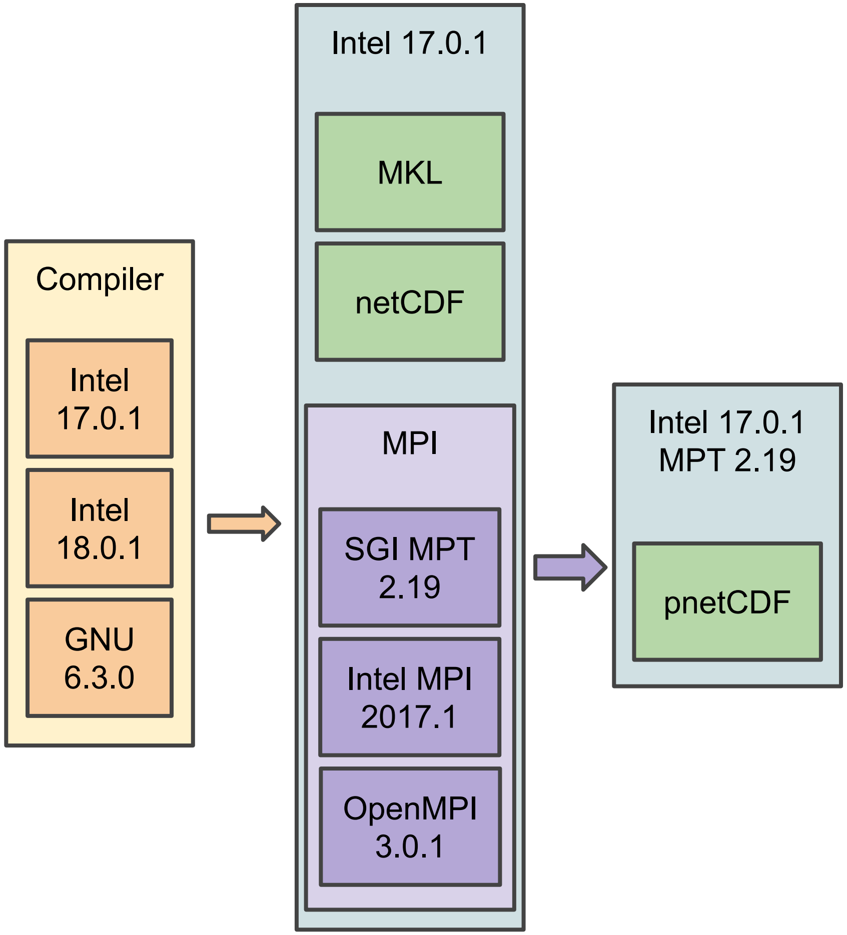
- For short transfers use **scp/sftp** to transfer files
- For large transfers use **Globus**
  - To use Globus, create a Globus ID if you need an account, and search for **NCAR GLADE** or **NCAR Campaign Storage** endpoints
  - CISL endpoints currently can be activated for up to 30-days
  - Globus has a web interface and a command-line interface
  - **Globus Connect Personal** can manage transfers from your local workstation as well

Environment Modules

- CISL installed software is provided as modules
- Modules provide access to runnable applications (compilers, debuggers, ...) as well as libraries (NetCDF, MPI, ...)
- Modules prevent loading incompatible software into your environment
- **Note that Cheyenne and Casper each have independent collections of modules!**

Using modules

- **module load/unload <software>**
- **module avail** - show all currently-loadable modules
- **module list** - show loaded modules
- **module purge** - remove all loaded modules
- **module save/restore <name>** create/load a saved set of software
- **module spider <software>** search for a particular module



Changing your default modules

- If you commonly load certain modules, you may wish to have them load automatically when logging onto a cluster
- The right way to do so is with saved module sets:

```
module load ncl python nco mkl
module save default
```

- Make multiple sets and load them using **module restore <set>**
- **Don’t put module load commands in your shell startup files!**

Available Software

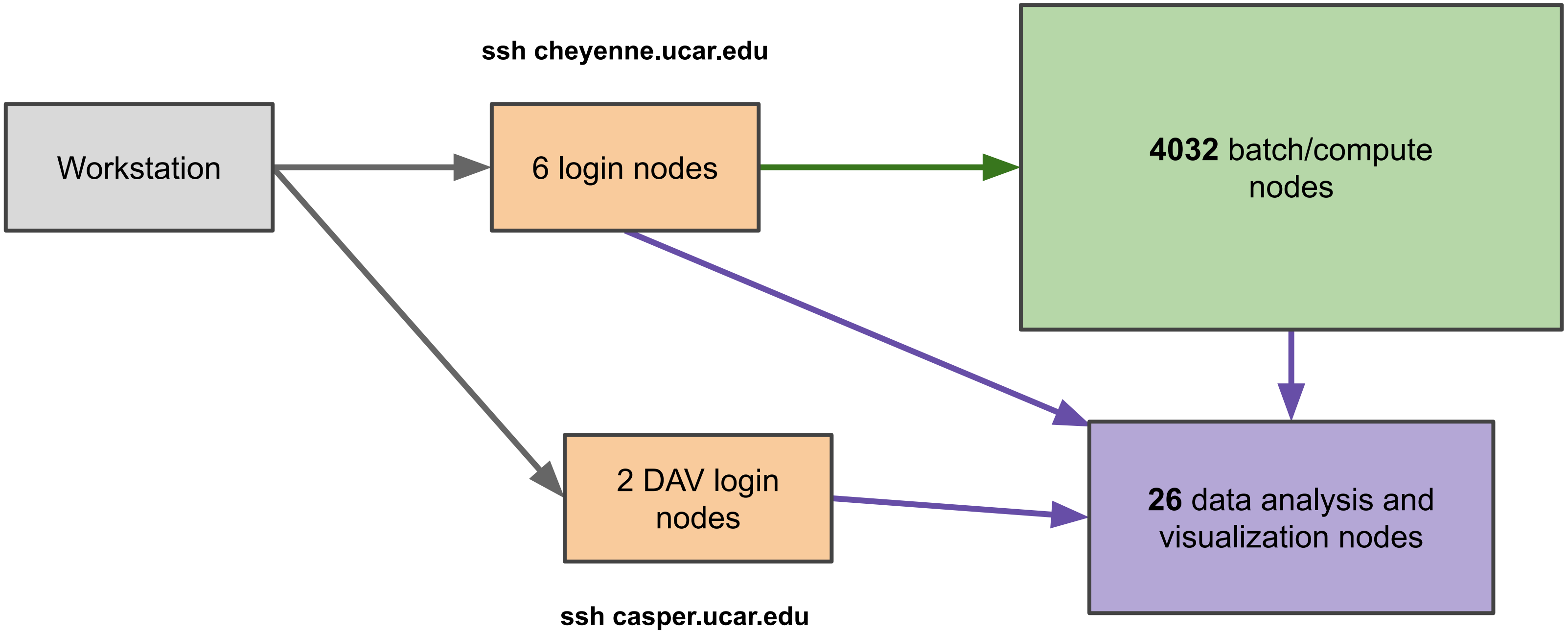
- Compilers (Intel, GNU, PGI)
- Debuggers / Performance Tools (DDT, MAP)
- MPI Libraries (MPT, Intel MPI, OpenMPI)
- IO Libraries (NetCDF, PNetCDF, HDF5)
- Analysis Languages (Python, Julia, R, IDL, Matlab)
- Convenience Tools (ncarcompilers, parallel, rclone)
- Many more ...

Considerations when compiling software

- Use **ncarcompilers** module along with library modules (e.g., netcdf) to simplify compiling and linking (*it adds include and link flags for you*)
- When using MPI, make sure you run with the same library with which you compiled your code
- Cheyenne and Casper have different CPUs and operating systems **We strongly recommend you build code on the machine on which you will run**

Use batch node jobs for large compute tasks

- Most tasks require too many resources to run on a login node
- Schedule these tasks to run on the Cheyenne compute nodes using the **PBS** batch system



ssh cheyenne.ucar.edu

Workstation

6 login nodes

4032 batch/compute nodes

2 DAV login nodes

26 data analysis and visualization nodes

ssh casper.ucar.edu

Cheyenne and Casper use separate allocations!

Use batch node jobs for large compute tasks

- Most tasks require too many resources to run on a login node
- Schedule these tasks to run on the Cheyenne compute nodes using the **PBS** batch system
- Jobs request a given number of compute tasks for an estimated wall-time on specified hardware
- Jobs use core-hours, which are charged against your selected project/account
  - Remaining resources are viewable in SAM
- Temporary files are often written by programs - set TMPDIR variable to scratch space to avoid job failures

Example PBS job scripts

Cheyenne

```
$ cat basic_mpi.pbs
#!/bin/tcsh
#PBS -N hello_pbs
#PBS -A <project_code>
#PBS -j oe
#PBS -o pbsjob.log
#PBS -q regular
#PBS -l walltime=00:05:00
#PBS -l select=2:ncpus=36:mpiprocs=36

### Set temp to scratch
setenv TMPDIR /glade/scratch/${USER}/temp
mkdir -p $TMPDIR

module load mpt/2.25

### Run MPT MPI Program
mpiexec_mpt ./hello_world
```

Casper

```
$ cat array_job.pbs
#!/bin/bash -l
#PBS -N job_array
#PBS -A project_code
#PBS -l select=1:ncpus=1:mem=4GB
#PBS -l walltime=00:10:00
#PBS -q casper
#PBS -J 2010-2020
#PBS -j oe

### Set temp to scratch
export TMPDIR=/glade/scratch/$USER/temp
mkdir -p $TMPDIR

module load mpt/2.25

### Run Array jobs program
./executable_name
data.year-$PBS_ARRAY_INDEX
```

Interacting with the PBS job scheduler

qsub <script> - submit batch job

qstat <jobid> - query job status

qdel <jobid> - delete/kill a job

qinteractive -A <project>

Run an interactive job

qcmd -A <project> -- cmd.exe

Run cmd.exe on a single compute node

Using OpenMP parallelism on Cheyenne

OpenMP Only

```
#!/bin/tcsh
#PBS -l select=1:ncpus=10:ompthreads=10

# Run program with 10 threads
./executable_name
```

Hybrid MPI/OpenMP

```
#!/bin/tcsh
#PBS -l select=2:ncpus=36:mpiprocs=12:ompthreads=3

module load mpt/2.19

# Run program with one MPI task and 36 OpenMP
# threads per node (two nodes)
mpiexec_mpt omplace ./executable_name
```

Using command file jobs on multiple data

```
./cmd1.exe < input1 > output1
./cmd2.exe < input2 > output2
./cmd3.exe < input3 > output3
./cmd4.exe < input4 > output4
```

cmdfile contents

```
#!/bin/tcsh
#PBS -l select=1:ncpus=4:mpiprocs=4

module load mpt/2.19

# This setting is required to use command files
setenv MPI_SHEPHERD true

mpiexec_mpt launch_cf.sh cmdfile
```

PBS Job script

Optimal if commands have similar runtimes

Requesting Specific Resources for Casper jobs

```
cat gpu_job.pbs
#!/bin/bash -l
#PBS -N mpi_job
#PBS -A project_code
#PBS -l
select=1:ncpus=4:mpiprocs=4:ngpus=4:mem=40GB
#PBS -l gpu_type=v100
#PBS -l walltime=01:00:00
#PBS -q casper
#PBS -j oe

export TMPDIR=/glade/scratch/$USER/temp
mkdir -p $TMPDIR

### Provide CUDA runtime libraries
module load cuda

### Run program
mpirun ./gpu_code.exe
```

- This job can only run on a node with 40 GB of free memory and 4 V100 GPUs
- If multiple resources are specified, they must be compatible, otherwise, the job will be stuck in a pending state

PBS queues on Cheyenne

PBS Queue	Priority	Wall clock	Details
premium	1	12 h	Jobs are charged at 150% of regular rate
regular	2	12 h	Most production compute jobs go here
economy	3	12 h	Jobs are charged at 70% of regular rate
share	N/A	6 h	Memory is shared among all users on a node Jobs are limited to 18 cores or less

Job charges depend on the queue:

Exclusive: wall-clock hours ✖ nodes ✖ 36 cores/node ✖ queue factor

Shared: core-seconds / 3600 (DAV jobs are shared as well)

Shell startup files - customizing your default environment

tcsh/csh

```
$ cat ~/.tcshrc
alias rm ”rm -i”

# Add programs built for each cluster
if ( $HOSTNAME =~ cheyenne* ) then
   setenv PATH ~/local/ch/bin:$PATH
else
   setenv PATH ~/local/dav/bin:$PATH
endif

# Settings for interactive shells
if ( $?prompt ) then
   set prompt = "%n@%m:%~"
endif
```

bash

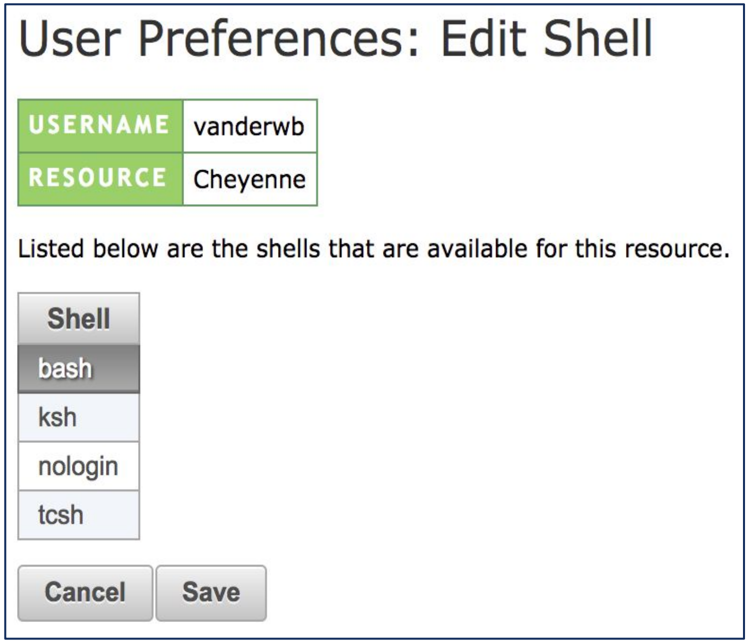
```
$ cat ~/.profile
alias rm=”rm -i”

# Add programs built for each
cluster
if [[ $HOSTNAME == cheyenne* ]];
then
   export PATH=~/local/ch/bin:$PATH
else
   export PATH=~/local/dav/bin:$PATH
fi

# Settings for interactive shells
if [[ $- == *i* ]]; then
   PS1="\u@\h:\w> "
fi
```

SAM (Systems Accounting Manager)

- Web access: https://sam.ucar.edu
- Log in with Duo authentication
- Can change some user settings (default shell, etc)
- Get information about available projects and remaining allocation balance
- See history of jobs and charges



User Preferences: Edit Shell

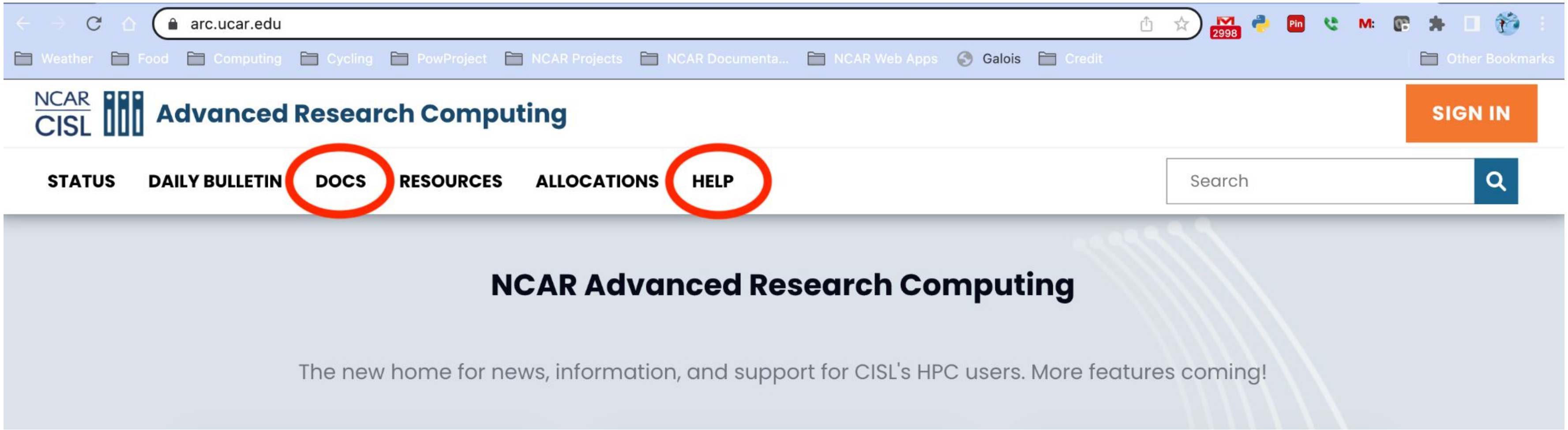
USERNAME	vanderwb
RESOURCE	Cheyenne

Listed below are the shells that are available for this resource.

Shell
bash
ksh
nologin
tcsh

Getting Help

https://arc.ucar.edu/



arc.ucar.edu

NCAR Advanced Research Computing

The new home for news, information, and support for CISL's HPC users. More features coming!

- Searchable Documentation
- Support Tickets with HPC Consultants