Software Tools for Scientific Research

Nathan Dunn
ndunn@cas.uoregon.edu

Craig Rasmussen
rasmus@cas.uoregon.edu

Kristi Potter
kpotter@uoregon.edu
Overview

• Signup
• Todos
• Questions?
• Finish Unix Commands
• ACISS
• Cloud Computing
• Practical Scripting?
Signup

- Signup going around:
  - if you’ve not sent me email, signup
Todo

- Signup form going around
- Get a laptop with UO network access
- Working Linux / Unix Environment
- [https://aciss.uoregon.edu/newuser](https://aciss.uoregon.edu/newuser)
- Add note / email if you want Cloud: Robert Yelle ryelle@uoregon.edu
Remote Copy: scp

• scp <filename> <user>@<host>::<directory>

• copy files TO home directory on aciss

NathanDunn:unix-biologists% scp iris_ids* ndunn@aciss.uoregon.edu:
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
iris_ids_sorted.txt 100% 11MB 10.9MB/s 00:01

• copy files TO “/tmp” on aciss

NathanDunn:unix-biologists% scp iris_ids* ndunn@aciss.uoregon.edu:/tmp
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
iris_ids_sorted.txt 100% 11MB 10.9MB/s 00:01
Remote Copy: scp

- scp <user>@<host>:<directory / file> <location / filename>

- copy files FROM home directory on aciss

```
NathanDunn@unix-biologists$ scp ndunn@aciss.uoregon.edu:iris_ids.txt .
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
```

- copy files FROM “/tmp” on aciss

```
NathanDunn@unix-biologists$ scp ndunn@aciss.uoregon.edu:/tmp/iris_ids.txt .
ndunn@aciss.uoregon.edu's password:
iris_ids.txt 100% 11MB 10.9MB/s 00:01
```
Remote Copy: sftp

- `sftp <user>@<host>`
- similar but interactive

```bash
NathanDunn@unix-biologists% sftp ndunn@aciss.uoregon.edu
ndunn@aciss.uoregon.edu's password:
Connected to aciss.uoregon.edu.
sftp> ls
Downloads  blast_test.e100277  blast_test.e100924  blast_test.e100925
blast_test.o100277  blast_test.o100924  blast_test.o100925  hg
iris_ids.txt     iris_ids_sorted.txt  old.local
sftp> put iris_ids*
Uploading iris_ids.txt to /ibrix/home2/ndunn/iris_ids.txt
iris_ids.txt     100%  11MB  10.9MB/s  00:01
Uploading iris_ids_sorted.txt to /ibrix/home2/ndunn/iris_ids_sorted.txt
iris_ids_sorted.txt  100%  11MB  10.9MB/s  00:01
```

Monday, July 8, 13
Remote Copy: wget

• downloads files + web pages
• wget <url>

  => `allcontig.agp.gz.1'
Resolving ftp.ncbi.nih.gov... 130.14.250.11
Logging in as anonymous... Logged in!
  ==> SYST ... done.  ==> PWD ... done.
  ==> TYPE I ... done.  ==> CWD (1) /genomes/D_rerio ... done.
  ==> SIZE allcontig.agp.gz ... 1486337
  ==> PASV ... done.  ==> RETR allcontig.agp.gz ... done.
Length: 1486337 (1.4M) (unauthoritative)

100%[===============================================] 1,486,337 197K/s in 6.3s

2012-05-14 15:20:47 (231 KB/s) - `allcontig.agp.gz.1' saved [1486337]
Remote Copy: curl

- downloads files
- `curl -o <output> <url>

```
% Total    % Received % Xferd Average Speed Time   Time  Time  Current
     0     0   181k      0   --:--:--     0        0           0          0          0          0          0
100   1451k  100 1451k  0  0  197k  0  0:00:07  0:00:07  --:--:--  --:--:--
NathanDunn@unix-biologists% ```
Advanced Search: 
mdfind

- mac only ... but fast ... can search in
- mdfind -name <pattern>
- mdfind <pattern>
- mdfind -name .key | grep key$

NathanDunn@casspr% mdfind -name ".key" | grep key$
/Users/NathanDunn/Library/Preferences/IntelliJIdea11/idea11.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week1-presentations.key
/Users/NathanDunn/hg/casspr/presentations/uogenesis/uogen_present.key
/Users/NathanDunn/hg/casspr/presentations/unix-biologists/week3-presentation.key

Monday, July 8, 13
Advanced Search: locate

• Similar to mdfind, but on Linux
• Will need to initialize database if not done
Advanced Search

• sed: stream editor

• sed 's/\|/,/g' output/phenotypes.txt > output/phenotypes.csv

• awk
What is ACISS?

- High Performance Computer:
  - 128 Basic Nodes: 12 cores, 72 GB memory
  - 52 GPU Nodes: Same as basic, plus 3 Nvidia 2070 GPUs
  - 16 Fat Nodes: 32 cores, 384 GB memory
  - 400 TB Usable storage
What is ACISS?

- Both Cluster and Cloud
- For researchers, students, and collaborators
- Free
Cluster

- Shared super-computer
- Configured with job scheduler (qsub)

```
aciss.uoregon.edu
```

```
\ssh
```

```
\\%
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
\activity
```

```
```
Cloud

- Personal super-computer(s)
- Create instances (booted machine) as necessary
- Examples: Amazon and ACISS
- Local Example: Virtual Box (demo)
Cloud

Inactive

Image

Snapshot

Active

Instance(s)

Boot / Launch

Take Snapshot(s)

~ Close Laptop

Relaunch

ssh

web
# Cloud vs Cluster

## Cloud
- Isolated environment
- No timeline
- Quota on size
- Better for service

## Cluster
- Shared environment
- Closer to hardware
- Scheduled
- 1 day, 4 days, 2 week
- Better for batch jobs
Using ACISS Cloud

- Account
- Managing Instances in Cloud
- Access to ACISS Cluster
- Run programs
- Move Files
- Install programs
- Large Instances
Cloud Account

- [http://aciss.uoregon.edu/newuser](http://aciss.uoregon.edu/newuser) (ask at signup)
- or - Email for OpenStack account: [ryelle@cas.uoregon.edu](mailto:ryelle@cas.uoregon.edu)

Resources:

- [http://blogs.uoregon.edu/casspr/2012/08/03/create-a-new-cloud-instance/](http://blogs.uoregon.edu/casspr/2012/08/03/create-a-new-cloud-instance/)
- [http://blogs.uoregon.edu/casspr/2013/07/08/cloud-slides-for-advanced-users/](http://blogs.uoregon.edu/casspr/2013/07/08/cloud-slides-for-advanced-users/)
- [http://aciss.uoregon.edu/wiki/Cloud_Deployment](http://aciss.uoregon.edu/wiki/Cloud_Deployment)
Cloud ACISS Interface

- [https://openstack.aciss.uoregon.edu](https://openstack.aciss.uoregon.edu) (demo)
  - overview
  - associate key
  - start instance from an image
  - login to instance
  - save snapshot
  - relaunch snapshot
Float IP’s

ACISS

Internal IP

Float IP

Internal IP

World

<table>
<thead>
<tr>
<th>Instance Name</th>
<th>IP Address</th>
<th>Size</th>
<th>Keypair</th>
<th>Status</th>
<th>Task</th>
<th>Power State</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech recognition -</td>
<td>10.0.0.32 128.223</td>
<td>m1.large</td>
<td>16GB RAM</td>
<td>4 VCPU</td>
<td>20GB Disk</td>
<td>-</td>
<td>Active None Running</td>
</tr>
<tr>
<td>large</td>
<td>224.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>qiime-server</td>
<td>10.0.0.17 128.223</td>
<td>m1.large</td>
<td>16GB RAM</td>
<td>4 VCPU</td>
<td>20GB Disk</td>
<td>-</td>
<td>Active None Running</td>
</tr>
<tr>
<td></td>
<td>224.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metagenomics-demo</td>
<td>10.0.0.19 128.223</td>
<td>m1.small</td>
<td>4GB RAM</td>
<td>1 VCPU</td>
<td>20GB Disk</td>
<td>precog</td>
<td>Active None Running</td>
</tr>
<tr>
<td></td>
<td>224.29</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Access to Cloud

• As Root
  • chmod 0400 ~/.ssh/key.pem
  • ssh -i ~/.ssh/key.pem root@128.223.225.5

• As User
  • adduser <user>  # as root
  • ssh <user>@128.223.225.5
Running in Cloud

- No different than running locally
- blastn -query OTTDART00000050916.fa -db ~/aciss/sequences/GenBank/blast/db/refseq_rna -out blast_cloud.txt -num_threads 4
Files to Cloud

- scp <file> -i ~/.ssh/key.pem root@128.225.223.5:
- scp <file> <user>@128.225.223.5:

- Mount Directory (e.g., sshfs)
  - mkdir ~/aciss/sequences
  - sshfs <user>@aciss.uoregon.edu:/research/sequences /home/ndunn/aciss/sequences
Installing on Cloud

• Same as regular computer

• Ubuntu (as root or using sudo)
  • apt-get install <package>
  • e.g., apt-get install ncbi-blast+ sshfs

• RedHat (as root or using sudo)
  • yum install <package>
  • e.g., yum install ncbi-blast+ sshfs
Large Instances

• All methods: http://blogs.uoregon.edu/casspr/2012/09/21/large-instances-on-the-aciss-cloud/

• Can use sshfs

• Need fast rebootable disk (<1 TB):
  • http://blogs.uoregon.edu/casspr/2012/06/27/aciss-cloud-attaching-a-volume-to-an-instance/
Long Running Instances

- `nohup command &`
- Writes to nohup.out where run
- Can ssh back in and still running
- `tail -f nohup.out`
Long Running Instances

- `screen`
  - opens a “new terminal”
  - run command as above
  - “Ctrl-A” “d” (disconnects) .. exit
- Reconnect:
  - `screen -ls`
  - `screen -r process`
Using ACISS Cluster

- Account
- Connect
- Run programs
- Install programs
- Move files
- Large Instances
Cluster Account

• ACISS available to researchers, students, and collaborators.

• https://aciss.uoregon.edu/newuser

• Ask for cloud if needed (or email ryelle@cas.uoregon.edu)

• Resource: http://aciss.uoregon.edu
Connect to Cluster

- ssh <user>@aciss.uoregon.edu
Run Interactive

• Get onto node: qsub -q generic -I

  • time blastn -query OTTDART00000050916.fa -db /research/sequences/GenBank/blast/db/refseq_rna -out blast1.txt

  • time blastn -query OTTDART00000050916.fa -db /research/sequences/GenBank/blast/db/refseq_rna -out blast1.txt -num_threads 12
Run Scheduled Jobs

- Schedule on node
  - setup keys: [http://aciss.uoregon.edu/wiki/First_Time_Users](http://aciss.uoregon.edu/wiki/First_Time_Users)
  - `qsub small_test.sh`

Queues
- `qstat`
Run Scheduled Jobs

- small_test.sh

```bash
#!/bin/sh -l
#PBS -N job_name
#PBS -q generic
#PBS -l nodes=1:ppn=4
#PBS -d /home2/ndunn

module load blast

blastn
--query OTTDART00000050916.fa
--db /research/sequences/GenBank/blast/db/refseq_rna
--out .txt
--num_threads 4
```
Installing on Cluster

- module
- module avail
- module load <module>
- email Robert Yelle
  - ryelle@uoregon.edu
Move Files to Cluster

- scp <file> <user>@aciss.uoregon.edu:
- cyerbduck
- WinSCP
Why is my code slow?

- Writing to home directory (/home) on node
  - write to anything but /home
- Not actually using multiple nodes:
  - top
Summary

- Work on your TODO list!
- Get a working Linux / UNIX environment
- Review older class files
- Bring in / email us your (technical) problems
- Next time - Bash scripts
Select a month to query its usage:

<table>
<thead>
<tr>
<th>Month</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>April</td>
<td>2012</td>
</tr>
</tbody>
</table>

Active Instances: 3
Active Memory: 20GB
This Month’s VCPU-Hours: 627.73
This Month’s GB-Hours: 38639.51

Usage Summary

<table>
<thead>
<tr>
<th>Instance Name</th>
<th>VCPUs</th>
<th>Disk</th>
<th>RAM</th>
<th>Uptime</th>
</tr>
</thead>
<tbody>
<tr>
<td>blast test</td>
<td>2</td>
<td>50</td>
<td>4GB</td>
<td>1 week, 2 days</td>
</tr>
<tr>
<td>tiny</td>
<td>1</td>
<td>-</td>
<td>512MB</td>
<td>1 week, 2 days</td>
</tr>
<tr>
<td>clovr2</td>
<td>8</td>
<td>170</td>
<td>16GB</td>
<td>6 days, 14 hours</td>
</tr>
</tbody>
</table>

Displaying 3 items
Launch Instances

Server Name

User Data

Description:
Specify the details for launching an instance. The chart below shows the resources used by this project in relation to the project's quotas.

Project Quotas
Instance Count (7) 3 Available
VCPUs (12) 8 Available
Disk (280 GB) 720 GB Available
Memory (23040 MB) 28160 MB Available

Flavor
m1.tiny (1CPU / 0GB Disk / 512MB Ram)

Keypair
Select a keypair

Instance Count
1

Security Groups
- default
- tecsec

Boot From Volume

Cancel Launch Instance
### Floating IPs

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Instance</th>
<th>Floating IP Pool</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No items to display.</td>
<td></td>
</tr>
</tbody>
</table>

Displaying 0 items

### Security Groups

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>default</td>
<td>default</td>
<td></td>
</tr>
</tbody>
</table>

Displaying 1 item

### Keypairs

<table>
<thead>
<tr>
<th>Keypair Name</th>
<th>Fingerprint</th>
<th>Actions</th>
</tr>
</thead>
</table>

Displaying 1 item
<table>
<thead>
<tr>
<th>Image Name</th>
<th>Type</th>
<th>Status</th>
<th>Public</th>
<th>Container Format</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubuntu Server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>ubuntu-initrd.img-3.0.0-12-server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>ubuntu-vmlinuz-3.0.0-12-server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>Matlab Scientific Linux 6.2</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>Clovr Server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>clovr-initrd.img-2.6.32-21-server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>clovr-vmlinuz-2.6.32-21-server</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>Scientific Linux 6.2 (devel)</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>vmlinuz-2.6.32-220.7.1.el6.x86_64</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
<tr>
<td>initramfs-2.6.32-220.7.1.el6.x86_64.img</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>BARE</td>
<td>Launch</td>
</tr>
</tbody>
</table>
### Instances

<table>
<thead>
<tr>
<th>Instance Name</th>
<th>IP Address</th>
<th>Size</th>
<th>Status</th>
<th>Task</th>
<th>Power State</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>matlab test</td>
<td>128.223.225.6</td>
<td>2GB RAM</td>
<td>1 VCPU</td>
<td>10GB Disk</td>
<td>% Build</td>
<td>None</td>
</tr>
<tr>
<td>clovr2</td>
<td>128.223.225.5</td>
<td>16GB RAM</td>
<td>8 VCPU</td>
<td>10GB Disk</td>
<td>Active</td>
<td>None</td>
</tr>
<tr>
<td>nathan clovr server</td>
<td>128.223.225.13</td>
<td>4GB RAM</td>
<td>2 VCPU</td>
<td>10GB Disk</td>
<td>Error</td>
<td>None</td>
</tr>
<tr>
<td>tiny</td>
<td>128.223.225.13</td>
<td>512MB RAM</td>
<td>1 VCPU</td>
<td>0 Disk</td>
<td>Suspended</td>
<td>None</td>
</tr>
<tr>
<td>blast test</td>
<td>128.223.225.12</td>
<td>4GB RAM</td>
<td>2 VCPU</td>
<td>10GB Disk</td>
<td>Active</td>
<td>None</td>
</tr>
</tbody>
</table>

Displaying 5 items

### Volumes

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Size</th>
<th>Status</th>
<th>Attachments</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>disk1</td>
<td>-</td>
<td>10 GB</td>
<td>Available</td>
<td>-</td>
<td>Edit Attachments</td>
</tr>
<tr>
<td>galaxy VM</td>
<td>I will put the galaxy VM here. The minimum size should be 25 GB.</td>
<td>50 GB</td>
<td>Available</td>
<td>-</td>
<td>Edit Attachments</td>
</tr>
</tbody>
</table>

Displaying 2 items