Accessing a new super-computer

Mary A. Allen, 2021

The AWS (*Amazon Web Services)* supercomputer we used in class goes away Monday!!! AWS supercomputers cost, so it's nice if your campus has a cluster, supercomputer, or some high performance compute option.

Super computers on the Boulder/Anschutz campuses

Super computer name	**Run by**	**Who can get on**	**Security system**
fiji	BioFrontiers IT (bit-help@colorado.edu)	Boulder BioFrontiers faculty and their labs	You must be on campus or use vpn before login when off campus. Use your identikey and password as login information.
Summit	RC (Resource Computing)	Boulder campus - Anyone	Duo 2-factor Authentication
Rosalind	TICR High Performance Computing	CU Anschutz: Authorized faculty, students, and staff	Must be on campus or use vpn when off campus. Requires a password.
Bodhi	Dept of Biochemistry and Molecular Genetics, School of Medicine (david.farrell@cuanschutz.edu)	CU Anschutz Authorized faculty, students, and staff	Must be on campus or use VPN, and Duo 2-factor authentication using university credentials

Log in

Open terminal on a mac or a bash system on the pc like ubuntu. You will then login to the computer using your username (on fiji this is your identikey) and the machine name (example: fiji.colorado.edu).

Type$ ssh <username>@<computername>

The first time you log in it will ask you:

Are you sure you want to continue. Type$ yes

Super computers will either use a ssh key or a password (or both). If you type a password, you will see nothing. That's normal! It's a feature not a bug.

This is me logging onto fiji:

```
cu-biot-6-10:~ maryallen$ ssh allenma@fiji.colorado.edu
The authenticity of host 'fiji.colorado.edu (128.138.93.105)' can't be established.
ECDSA key fingerprint is SHA256:5HyiwNcswNas+K9fatox+9sJqNlAS9Q816/jHfCX8a4.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'fiji.colorado.edu' (ECDSA) to the list of known hosts.
allenma@fiji.colorado.edu's password:
```

This is after I log onto fiji.

```
Last login: Wed Jul 14 09:35:15 2021 from cu-biot-6-10.203.150.69.int.colorado.edu
+------------------------------------------------------------------------------+
|                    *****Warning Notice*****                                  |
|This system is restricted solely to University of Colorado authorized users  |
|for legitimate business purposes only.  The actual of attempted unauthorized |
|access, use or modification of this system is strictly prohibited by the      |
|University of Colorado.                                                       |
|The use of this system is monitored and recorded for administrative and       |
|security reasons.  Anyone accessing this system expressly consents to such    |
|monitoring and is advised that is such monitoring reveals possible evidence  |
|of criminal activity, University officals may provide evidence of such        |
|activity to law enforcement officials.  All users must comply with the        |
|University of Colorado security instructions regarding the protection of the |
|University of Colorado's information.                                         |
|                                                                              |
|Cluster usage information is available at the BioFrontiers IT computing page |
|https://bit.colorado.edu/biofrontiers-computing/fiji/                         |
                                    |
|                                                                              |
|For assistance, please submit a ticket to our ticketing system --             |
|bit-help@colorado.edu .                                                       |
|                                                                              |
+------------------------------------------------------------------------------+
```

```
-bash-4.2$ hostname
fiji-2.colorado.edu
-bash-4.2$ logout
```

To confirm you are on the super computer
Type$ hostname
And the computer will tell you its name.

This is my personal computer:

```
cu-biot-6-10:~ maryallen$ hostname
cu-biot-6-10.203.150.69.int.colorado.edu
```

Log out

Type$ logout

I’m on a new supercomputer. What do I need to know?

This section assumes the super computer uses slurm as a queue system. There are other possibilities including Torque and Moab. Ask the admin of the computer how jobs are scheduled/managed.

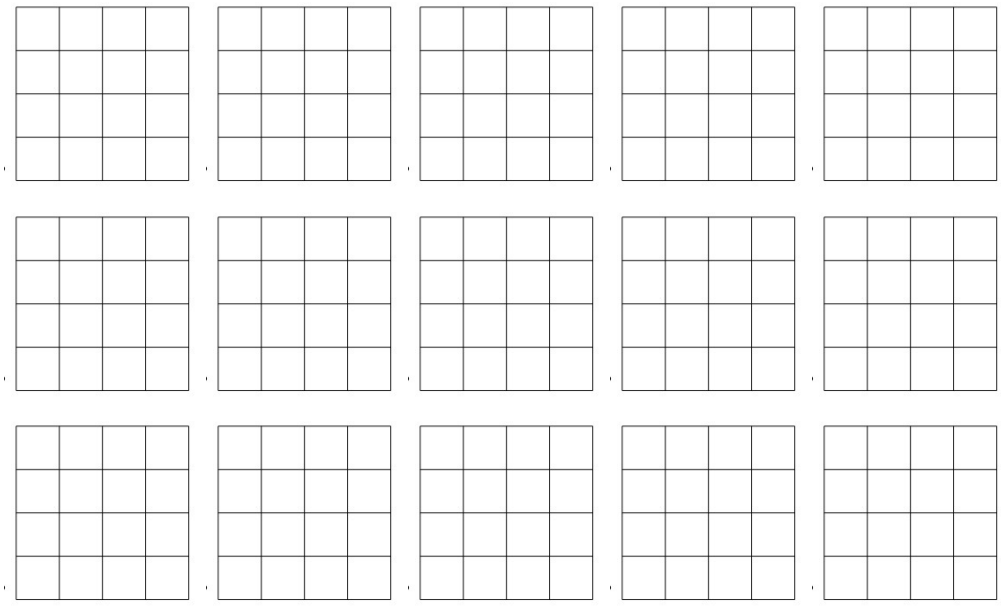
What are the partition names?

If the supercomputer has distinct resources (is heterogenous) then likely there will be different partitions available (i.e. what you do for the -p flag in a sbatch script). For a Slurm you can inquire about these options:

Type$ sinfo

How many Nodes and CPUs are there?

In the diagram below, a 4x4 box is a node (a single computer), smaller boxes inside of a node are CPUs (also known as ntasks or processors). The default for most programs is to assume they are running on a single node and 1 CPU. Writing software that can leverage multiple ntasks, processors, or CPUs is not hard, so check if the program you are using can take advantage of parallelization in this manner. Remember: The more CPUs the faster the program will run. Writing software to take advantage of multiple nodes is far more challenging and very few bioinformatics programs use multiple nodes, but it is possible! If you run into this situation, talk to your IT support guys for help in how to do this.



But how many nodes and CPUs does my cluster have?
Type$ scontrol show node

#Generally, you should use about half a node to run hisat2 and samtools sort.

How much is this going to cost?

- Storage of data costs! (disk drives, back ups, administration)
- Compute (running the programs) costs! (air conditioning, hardware, administration)
- Budget for the analysis (not just the library prep and the sequencing): it takes ***somebody*** time to do this analysis!
- Your best storage space estimate is 3-10x the size of your initial fastq files. This assumes you are not keeping intermediate files.

Where should I be putting data I'm working on vs. data I'm storing?

ASK, generally there is "local disk" (fast and physically nearby) and "network disk" (slower and possibly physically distant). Speed of the disk influences how fast your jobs can run. For example -- on fiji the scratch directories are the fastest.

ASK what is backed up?
Not all directories can be backed up. Back-ups are expensive (more disks, disks have to be setup and maintained, ideally are physically distant). On Fiji: home directories are backed up but scratch is not.

Generally: local disk i/o is much faster than network disk

If I'm not using (NOT SCRATCH)

And I don't have much space!
And I have to share space with my neighbors!
But I'm working here.

(SCRATCH)
Faster, easier on the computer when doing a lot of I/O

(NOT SCRATCH)
Slower I/O but generally cheaper

The other way in which these disk resources vary is in **total available space**. Generally speaking, faster disks are more expensive to buy initially -- and therefore there may be less of it available. Many computer systems have quotas (maximum amount of space available to you in either scratch or home) -- so be aware of these limits. Contact the administrative staff if space constraints are limiting -- as there may (or may not) be other options and/or workarounds available when very large amounts of disk are needed.

How much space am I using?

Type du -h