An introduction to the Rollins HPC cluster

Thomas Hsiao

September 29, 2022

Emory University

Table of contents

1. What is the Rollins HPC?

2. Using the RSPH HPC

3. Conclusion

What is the Rollins HPC?

Quick overview from the BIOS at RSPH website

”The RSPH HPC cluster is a system that consists of 25 compute nodes, 24 of which have 32 compute cores and 192GB of RAM each. The last node is a “large memory node” with 1.5 TB of RAM. These systems are connected together via 25GB Ethernet network, and all have access to a shared 1 Petabyte Panasas parallel file system.”

”Job scheduling is handled by the SLURM job scheduler, which is an application that currently runs on the majority of the Top 500 supercomputing sites in the world.”

Essentially - it’s a group of multiple computers with large number of resources that can talk to each other.

Should I be using the HPC?

If your code is taking hours or days to run, consider the HPC.

If you keep running out of memory (RAM), consider the HPC.

What scenarios would I use the HPC?

Standard examples include

1. Simulation studies with different parameters
2. Producing large numbers of plots
3. Running a large number of models
4. Running a big model like Stan or TMB

In general, if running multiple tasks that don’t depend on each other, HPC is a good option.

When is the HPC not the answer?

If you have one long operation that is not able to be split up.

If you’re just lazy and not writing your code in a smart way.

Using the RSPH HPC

How do I gain access?

Talk to Howard first. You need a faculty sponsorship to gain access.

Email help@sph.emory.edu requesting access while CC’ing Howard (or Howard emails directly?). Include your Emory NetID.

Logging in to the RSPH HPC

For Mac/Linux users use

```
ssh [NETID]@clogin01.sph.emory.edu
```

Where [NETID] is replaced with your unique Emory NetID.

Windows users can use the Linux terminal provided by WSL2, or use a terminal emulator like PuTTY.

How do I navigate the filesystem?

Review basic Linux commands (*`ls, cd, cp, mv, pwd, cat, mkdir, rm, etc.`*)

What filesystem should I be using?

As a member of SPENSER, you only need to focus on two directories.

/home/{NETID}: which is your home directory. This is where you start when you login to the cluster and only you have exclusive access.

/projects/hhchang: I would recommend putting all working directories on big projects here. You will need to request access to the Unix group spenser

Some terminology for parallel computing

Recall that the HPC is useful when you need to run multiple processes at the same time.

Each process is similar with slightly different parameter arguments.

Some terminology for parallel computing

There are many combinations to submit jobs in SLURM.

By far the most common way for us is to submit **job arrays**.

An example job array script

DEMO: Example code repo in Github repo

How to write a submit script

How to combine results

Conclusion

Resources

- SLURM Cheatsheet: lays out all the various ways to submit batch jobs.
- BIOS at RSPH HPC Guide: Good intro to cluster use with most of the basics.
- Sample SLURM code example on my GitHub
- Guide to using Rstudio on the cluster: requires Emory login
- Princeton SLURM guide: probably most in-depth and helpful resource