R on Supercomputers

Pramod Gupta

Department of Astronomy, University of Washington

Why use R on supercomputers?

- More memory e.g. 512GB RAM
- More disk space e.g. 100 TB
- More processors e.g. 1000s of CPUs
- Get your work done faster

Supercomputers are shared

- Users do not have administrator access
- Users do not do not have write access to default install paths
- Users request compute nodes from the Scheduler (e.g. Slurm, PBS)
- Users must have X11 software on their desktop/laptop to see plots interactively

Accessing pre-installed R

- Supercomputers use software modules
- module avail (show list of available modules)
- module load r3.4.1 (module name may be different)
- module list (show list of currently loaded modules)
- module unload r3.4.1

Installing R Centos 7/Redhat 7

- tar -xvf R-3.4.1.tar.gz
- cd R-3.4.1
- ./configure ––prefix=/disk1/mygroup/Rinstall
- make
- make install

Installing R on Centos 6/Redhat 6

- Problem: Centos 6/Redhat 6 have older versions of zlib, bzip etc.
- R 3.3 and later need more recent versions of zlib, bzip etc.
- Solution: Spack at https://github.com/LLNL/spack
- Spack builds all missing dependencies.

Installing R on Centos 6/Redhat 6

- git clone https://github.com/llnl/spack.git
- source spack/share/spack/setup-env.sh
- spack list (list available spack packages)
- spack info r (more information about the r package)
- spack install r@3.4.1 (installs R 3.4.1)

View R plots interactively

- Mac desktop/laptop: install XQuartz
- Windows desktop/laptop: install X11 software
- ssh -X myuserid@sc.xyzu.edu
- Get an interactive node from the scheduler
- module load r3.4.1 (module name may be different)
- Run R and make plots. Plots will show up on your desktop/laptop.

Slurm scheduler

- Get interactive node:
- srun -p mygroup −−time=2:00:00 −−mem=50G −−pty /bin/bash
- Submit a batch job:
- sbatch -p mygroup -A myaccount myscript.slurm

PBS/Torque scheduler

- Get an interactive node:
- qsub -I -V -l walltime=2:00:00
- Submit a batch job:
- qsub myscript.pbs

Compute nodes have many cores

- Problem: Compute nodes have many cores e.g. 12, 16, 28, ...
- How can we use all the cores?
- Solution: Parallel programming e.g.
- GNU parallel, R parallel package, Rmpi etc.
- Above list is in order of increasing complexity.

Use all cores with GNU parallel

- First make a file mylistofwork like below:
- Rscript file1.R
- Rscript file2.R
- ...
- Rscript file100.R
- Next use GNU parallel:
- module load r3.4.1
- cat mylistofwork | parallel

Conclusion

- Get an account on a neigborhood supercomputer
- Get access to more memory, disk and CPUs
- Get results faster

Questions?