Workflow Management Software on HPCC

CMSE 890-602

Why use High Performance Computing Clusters?

- Access to lots of computational resources
  - CPUs
  - GPUs
  - Memory
  - Storage
- High speed network connections
- Run workflow steps in parallel
- Accelerate individual workflow steps
- Much cheaper than commercial cloud resources
- National access via…ACCESS
  - https://access-ci.org

Job scheduling

- HPCCs require management of user computing
- Job schedulers control:
  - Resource usage (CPU, GPU, memory etc)
  - Time usage
  - Priority access
- Users submit jobs to the scheduler via command line or scripts
- Common schedulers: SLURM, Torque

High throughput computing

- Subtly different from regular HPCC use
- Running many thousands of small jobs at once, instead of a few jobs with many resources each
- Uses distributed resources across many HPCCs
- Jobs are often “pre-emptible” i.e. local users can kick you off
- Open Science Grid
  - https://osg-htc.org/

Workflow manager interface

- Automatic creation of jobs
- Allocate per-process/rule resources
- Both managers we have discussed (SnakeMake and Nextflow) use Executors for scheduling
- SnakeMake has somewhat better SLURM integration
  - Including HPCC profiles, default resource allocation
- Nextflow has better commercial cloud integration

In-class assignment

Go to https://msu-cmse-courses.github.io/CMSE_890-602_snakemake/

Scroll to section 3.16

We will go through this together!

Homework

Work on your semester project!