Advanced Scheduling

https://github.com/ResearchComputing/USGS_2016_02_09-10/

February 10, 2016
Timothy Brown

Research Computing
UNIVERSITY OF COLORADO **BOULDER**

SLURM

Simple Linux Utility for Resource Management

- Allocates resources (compute nodes).
- Framework for starting, executing, and monitoring work.
- Manages a queue for pending work.

User Commands

- `salloc` for job allocation
- `srun` for running a job (either within an allocation (job step), or as a single allocation).
- `scancel` canceling jobs.
- `squeue` querying the job queue.

All commands have man-pages (and a lot of options).

Array Jobs

- Submitting and managing collections of jobs quickly and easily.
  - A single `SLURM_ARRAY_JOBID`.
  - Each element having an individual `SLURM_JOBID` and `SLURM_ARRAY_TASK_ID`.

script

```bash
#!/bin/bash
#SBATCH --job-name test-array
#SBATCH --time 5:00
#SBATCH --nodes 1
#SBATCH --output example-array-%a.out

filename=data.${SLURM_ARRAY_TASK_ID}
echo "processing ${filename}"
./a.out ${filename}
```

```
yeti-login01 ~$ sbatch --array 0-5 ./script
...
 processing data.0
 processing data.1
 processing data.2
 processing data.3
 processing data.4
```

Job Dependencies

It is possible to specify job dependencies with `-d`.

- `after` clause is satisfied after all jobs specified have started.
- `afterany` clause is satisfied after the specified jobs have complete.
- `afterok` clause is satisfied after all the jobs have complete successfully.
- `afternotok` clause is satisfied after any of the jobs have complete with at least one not completing successfully.

Distribution of Tasks

$$Total\,cpus\,requested = (Nodes) \times (S \times C \times T)$$

- High Level `–B S[:C[:T]]`
  - `S` number of sockets per node to allocate
  - `C` number of cores per socket to allocate
  - `T` number of threads per core to allocate

- Distribution options `–m X:Y`
  where `X` is across nodes and `Y` is across sockets.

  Options for nodes (`X`).
  - `block` consecutive tasks share a node.
  - `cyclic` round-robin.
  - `plane` distribute blocks in a specified size (on-line documentation).

  Options for sockets (`Y`).
  - `block` consecutive tasks share a socket.
  - `cyclic` round-robin.
  - `fcyclic` round-robin, however tasks requiring more than one cpu will be allocated in a cyclic fashion.

MPMD

If a job needs a custom task list the environment variable `SLURM_TASKS_PER_NODE` can be altered.

script

```
#!/bin/bash
#SBATCH --job-name io_test
#SBATCH --time 5:00
#SBATCH --nodes 4
#SBATCH --ntasks-per-node 12

export SLURM_TASKS_PER_NODE='1,12(x2),6'
mpiexec ./a.out
```

Questions?

Online Survey

<Timothy.Brown-1@colorado.edu>

License