HiPerGator 2.0: Running Jobs & Submission Scripts

Matt Gitzendanner *magitz@ufl.edu*

HiPerGator 2.0
The University of Florida Supercomputer

UF | Information Technology

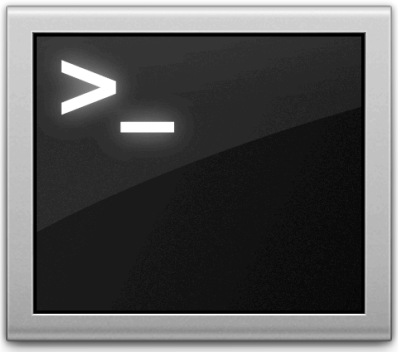
OneIT for the #GatorGood

HiPerGator 2.0
The University of Florida Supercomputer

UF | Information Technology

OneIT for the #GatorGood

Research Computing



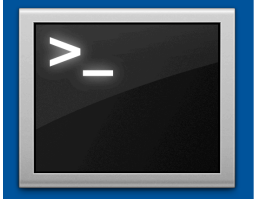
How do I get my jobs started?

UF | Information Technology

OneIT for the #GatorGood

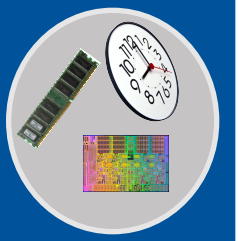
Cluster basics

User interaction



Login node (Head node)

Scheduler



Tell the scheduler what you want to do

Compute resources

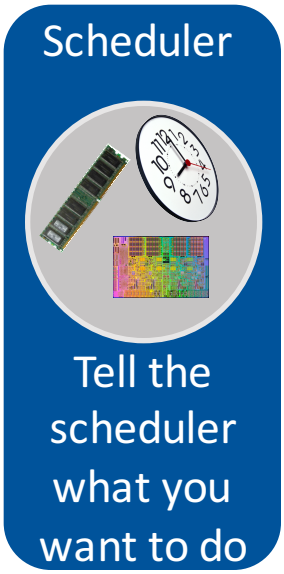
Your job runs on the cluster

UF | Information Technology

OneIT for the #GatorGood

Scheduling a job

- Need to tell scheduler what you want to do
  - **How many CPUs** you want and how you want them grouped
  - **How much RAM** your job will use
  - **How long** your job will run
  - The commands that will be run



Basic SLURM job script

```
#!/bin/bash
#SBATCH --job-name=test      #A name for your job
#SBATCH -o job_%j.out        #Name output file
#SBATCH --mail-type=ALL      #What emails you want
#SBATCH --mail-user=<Email address>    #Where
#SBATCH --ntasks=1           #Optional—single CPU
#SBATCH --mem-per-cpu=100mb  #Per core memory
#SBATCH -t=00:01:00      #Walltime in hh:mm:ss
                          #or d-hh:mm:ss

hostname
module load python
python -V
```

SLURM CPU Requests

- Nodes: `--nodes` or `-N`
  - Request a certain number of physical servers
- Tasks: `--ntasks` or `-n`
  - Total number of tasks job will use
- CPUs per task: `--cpus-per-task` or `-c`
  - Number of CPUs per task

HiPerGator 2.0 Compute Servers:
- 32 cores (2 X 16-core Intel Xeon CPUs)

SLURM CPU Requests

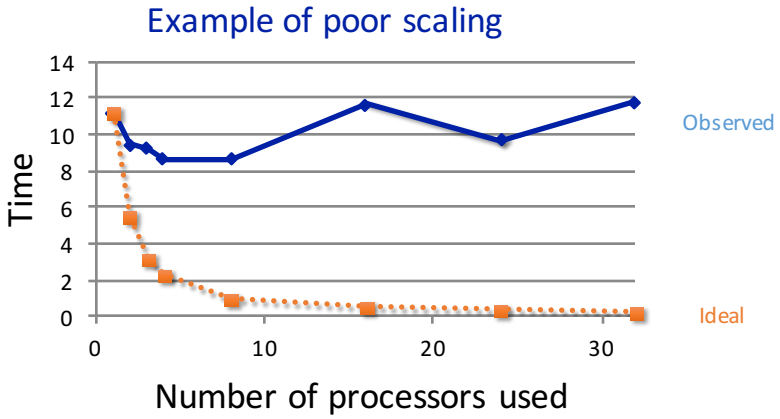
- For single processor jobs
  - `--ntasks=1` **(or omit)**
- For parallel jobs on a single node:
  - `--cpus-per-task=8`

SLURM CPU Requests

- For MPI jobs
  - `--ntasks=32`
  - Gets 32 cores for 32 MPI ranks
  - SLURM will determine node layout
- For Hybrid MPI/OpenMP jobs
  - `--ntasks=4` (4 MPI ranks)
  - `--cpus-per-task=8`
  - `(--nodes=4)` Not needed unless you really want 4 different nodes

Parallel efficiency

• How well does your application scale?



SLURM Memory Requests

- Memory: `--mem-per-cpu=1gb`
  - Can use mb or gb
  - No decimal values: use 1500mb, not 1.5gb

HiPerGator 2.0 Compute Servers:
- 128 GB total RAM (vs 256 GB on HPG1)
- Diskless servers: OS takes ~8GB RAM

SLURM Memory Requests

- Users commonly request 1,000s of times more memory than needed

Wasted RAM leads to idle CPUs and low job throughput

SLURM Time Request

- Time: `--time` or `-t`
  - 120 (minutes)
  - 2:00:00 (hh:mm:ss)
  - 7-0 (days-hours)
  - 7-00:00 (days-hh:mm)
  - 7-00:00:00 (days-hh:mm:ss)

SLURM output/error files

- `#SBATCH -o output.file`
- `#SBATCH -e error.file`
- `#SBATCH -o output.file` #W/o -e combined
- Can also use `--output` and `--error`

- `#SBATCH -o JobFile.%j.out`
  - Use %j instead of $SLURM_JOBID

SLURM

- Note that multi-letter directives are double-dash:
  - **`--mail-type`**
  - **`--ntasks`**
  - **`--mem-per-cpu`**

```
sbatch: error: distribution type
'ail-type=ALL' is not recognized
```

- Do not use spaces with =
  - **`--mail-user=magitz@ufl.edu`** ✔
  - **`--mail-user magitz@ufl.edu`** ✔
  - not: **`--mail-user= magitz@ufl.edu`**

Quality of Service (--qos)

- Each group has two QOS options
  - Investment QOS:
    - The NCUs the group has purchased
    - **`--qos=group`** (or leave off as this is default)
  - Burst QOS:
    - The burst capacity, available when idle resources are available on the cluster
    - **`--qos=group-b`**
- Users can choose higher priority, or larger pool of resources

SLURM Task Arrays

- `#SBATCH --array=1-200%10`
- Similar to Moab: range with % to limit number of jobs at a time
- `$SLURM_ARRAY_TASK_ID`
- Output file naming:
  - %A: job id
  - %a: task id
  - Output.%A_%a.out

SLURM environment

- SLURM inherits your environment
  - This includes present working directory
    - Don't need cd $SLURM_SUBMIT_DIR
  - Modules that are loaded
  - **Be careful of conflicting modules**

Emails

```
Job ID: 94392
Cluster: hipergator
User/Group: magitz/ufhpc
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 4
CPU Utilization: 00:00:44
CPU Efficiency: 52.38% of 00:01:24 core-walltime
Memory Utilization 1.52 MB
Memory Efficiency: 0.04% of 4.00 GB
```


Emails

```
Job ID: 5019
Cluster: hpg1
User/Group: magitz/ufhpc
State: CANCELLED (exit code 0)
Cores: 1
CPU Utilization: 00:00:00
CPU Efficiency: 0.00% of 00:00:00 core-walltime
Memory Utilization 1.26 MB
Memory Efficiency: 126.17% of 1.00 MB
```

Job error file:

```
slurmstepd: Job 5019 exceeded memory limit (1292 > 1024), being
killed
slurmstepd: Exceeded job memory limit
slurmstepd: *** JOB 5019 ON dev1 CANCELLED AT 2016-05-16T15:33:27
***
```

Development sessions

- Either:
  - **module load ufrc**
  - Followed by
    - **srundev**
    - **srundev -t 60:00**
- Or
  - **srun -p hpg2-dev --pty -u bash -i**
  - **srun -p hpg2-dev -t 60:00 --pty -u bash -i**

Checking on jobs

- squeue
- sacct
- See wiki.rc.ufl.edu/doc/SLURM_Commands
- See http://slurm.schedmd.com/

Example files

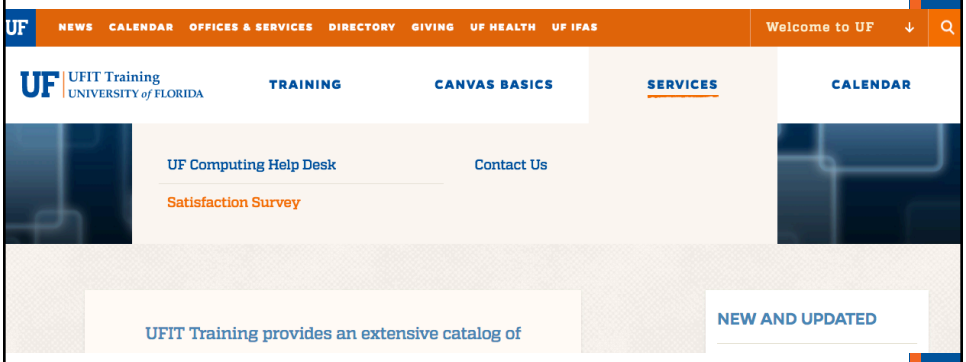
```
cd /ufrc/group/user/
mkdir SLURM_examples
cd SLURM_examples
cp /ufrc/data/training/SLURM/*.sbatch .
```

UF | Information Technology

OneIT for the #GatorGood

Satisfaction Survey

- training.it.ufl.edu



UF | Information Technology

OneIT for the #GatorGood

Next Week:

- Open Q&A session
  - 11:00am
  - NPB 2205

Support

- Support requests

- Web page and wiki