What is SLURM?

**Simple Linux Utility for Resource Management**

- Manages compute resources.
- Schedules jobs using those resources.
- Originally developed at Lawrence Livermore National Laboratory, but now maintained and supported by SchedMD.
- Open-source, GPL 2.0 licensing.
- Supports plugins for extending or enhancing functionality.
- Increasingly being used at academic research compute centers and national labs.
SLURM Features

Excellent performance
• Able to process tens of thousands of jobs per hour
• High throughput for smaller jobs (accepts up to 1,000 jobs per second)
• Fault tolerant

Supports Control Groups (cgroups)
• A cgroup is a Linux mechanism for aggregating and partitioning sets of tasks where resources are shared among multiple users
• Allows memory and CPU requests to be enforced on compute nodes

Uses a database to store user and account information, as well as job statistics
SLURM at ACCRE

scontrol
squeue
sinfo
scancel
rtracejob
stracejob
sacct
slurmdb
SLURM Database (mysql)

SLURM Controller (primary)
SLURM Controller (backup)

Compute Nodes

slurmdb
SLURM Database (mysql)
SLURM Commands

sbatch

Advanced Computing Center for Research and Education

www.accre.vanderbilt.edu

Follow us on twitter: @ACCREVandy

GitHub repos: https://github.com/accre
Submit Jobs with sbatch

1. Login to the cluster
   a. ssh <vunetid>@login.accre.vanderbilt.edu

2. Create a job script
   a. Use one of several editors: vi, emacs, or nano
   b. nano is likely best for novice Linux users
   c. Name the script in a meaningful way. Many users use “.slurm” as an extension to easily identify job scripts.
Submit Jobs with sbatch

d. Example script name:
   • monti_carlo_sym.slurm

3. Submit this script to the cluster with sbatch:
   a. sbatch monti_carlo_sym.slurm

4. Slurm will respond with a jobid:
   a. Submitted batch job 3115419
A Simple Job Script

#!/bin/bash

#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --mem=1G
#SBATCH --time=0-01:30:00
#SBATCH --output=my_output.txt
#SBATCH --mail-user=vunetid@vanderbilt.edu,you@gmail.com
#SBATCH --mail-type=ALL
#SBATCH --job-name="just_a_test"

# This will be run once for a single process
/bin/hostname
About the Simple Job Script

Requesting a single node: \texttt{--nodes=1}

Informing SLURM there will be just one task: \texttt{--ntasks=1}

Requesting one gigabyte of memory: \texttt{--mem=1G}

Requesting one hour and thirty minutes of time: \texttt{--time=0-01:30:00}

The format for time is \texttt{dd-hh:mm:ss} and the \textbf{maximum} amount of time is 14 days: \texttt{14-00:00:00} (the day specifier is optional).

Output not sent to a file goes here: \texttt{--output=my_output.txt}

Send email here: \texttt{--mail-user=vunetid@vanderbilt.edu,you@gmail.com}

Email when job starts, ends, or aborts: \texttt{--mail-type=ALL}

The name of the job: \texttt{--job-name="just_a_test"}

This will be run once as a single process: \texttt{/bin/hostname}
#!/bin/bash

#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=5
#SBATCH --mem=10G
#SBATCH --time=0-00:60:00
#SBATCH --output=my_output.txt
#SBATCH --mail-user=vunetid@vanderbilt.edu,you@gmail.com
#SBATCH --mail-type=ALL
#SBATCH --job-name="job_with_five_threads"

python multi threaded.py
About Multi-Threaded Job Script

Generally, multi-threaded jobs run on a single node, but use more than one processor on the node. The key request is “--cpus-per-task”. Also note that this job requests one hour of compute time. That could have been requested with “--time=01:00:00”

```
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=5
#SBATCH --mem=10G
#SBATCH --time=00:60:00
#SBATCH --output=my_output.txt
#SBATCH --mail-user=vunetid@vanderbilt.edu,you@gmail.com
#SBATCH --mail-type=ALL
#SBATCH --job-name="job_with_five_threads"
python multi_threaded.py
```
Overview

• Job arrays offer a mechanism for submitting and managing collections of similar jobs quickly and easily.

• Job arrays with many tasks can be submitted in milliseconds.

• All jobs must have the same initial options (e.g. size, time limit).

• Users may limit how many such jobs are running simultaneously.

• Job arrays are only supported for batch jobs.
Job ID and Environment Variables

Each array job has access to two environment variables

- **SLURM_ARRAY_JOB_ID**
  - will be set to the first job ID of the array.
- **SLURM_ARRAY_TASK_ID**
  - will be set to the job array index value.

Two additional options are available for file names.

- **%A**
  - will be replaced by the value of **SLURM_ARRAY_JOB_ID**
- **%a**
  - will be replaced by the value of **SLURM_ARRAY_TASK_ID**
#!/bin/bash

#SBATCH --mail-user=vunetid@vanderbilt.edu
#SBATCH --mail-type=ALL
#SBATCH --ntasks=1
#SBATCH --time=02:00:00
#SBATCH --mem=5G
#SBATCH --array=0-250
#SBATCH --output=ruby_array_job_slurm_%A_%a.out

echo "SLURM_JOBID: " $SLURM_JOBID

echo "SLURM_ARRAY_TASK_ID: " $SLURM_ARRAY_TASK_ID

echo "SLURM_ARRAY_JOB_ID: " $SLURM_ARRAY_JOB_ID

setpkgs -a ruby2.2.0

ruby matrix_transform_$(SLURM_ARRAY_TASK_ID).rb
**sbatch Commands**

<table>
<thead>
<tr>
<th>Command List</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>--nodes=[count]</td>
<td>Number of nodes requested</td>
</tr>
<tr>
<td>--tasks-per-node=[count]</td>
<td>Processes per node</td>
</tr>
<tr>
<td>--ntasks=[count]</td>
<td>Total number of processes</td>
</tr>
<tr>
<td>--cpus-per-task=[count]</td>
<td>Number of cores per process</td>
</tr>
<tr>
<td>--nodelist=[nodes]</td>
<td>Prefer jobs to run here</td>
</tr>
<tr>
<td>--exclude=[nodes]</td>
<td>Prefer jobs <strong>not</strong> to run here</td>
</tr>
<tr>
<td>--time=[min] or –time=days-hh:mm:ss</td>
<td>Wall time limit</td>
</tr>
<tr>
<td>--mem=[amount][M or G or T]</td>
<td>Memory per node (avoid --mem-per-cpu)</td>
</tr>
<tr>
<td>--mem-per-cpu=[amount][M or G or T]</td>
<td>Memory allocated per CPU (avoid --mem)</td>
</tr>
</tbody>
</table>
## sbatch Commands

<table>
<thead>
<tr>
<th>Command List</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>--output=[file_name]</td>
<td>STDOUT goes here</td>
</tr>
<tr>
<td>--error=[file_name]</td>
<td>STDERR goes here</td>
</tr>
<tr>
<td>--array=[array specification]</td>
<td>Job arrays</td>
</tr>
<tr>
<td>--array=[array specification]%[count]</td>
<td>Permit only “count” jobs to run at one time</td>
</tr>
<tr>
<td>--mail-user=[email_address]</td>
<td>Email notifications sent here</td>
</tr>
<tr>
<td>--email-type=[BEGIN</td>
<td>END</td>
</tr>
<tr>
<td>--account=[group name]</td>
<td>Run this job under this group</td>
</tr>
<tr>
<td>--job-name=[name]</td>
<td>Name this job</td>
</tr>
<tr>
<td>--contstrain=[attribute]</td>
<td>Request nodes of a certain type, e.g., “intel”</td>
</tr>
</tbody>
</table>
SLURM Commands

scancel
Delete Jobs with scancel

scancel <jobid>

- Jobs **not** running will be deleted
- Running jobs will be killed

- `scancel <jobid1> <jobid2> <jobid3>`

- Multiple jobs may be listed separated by spaces.
More on scancel

• If the job ID of a job array is specified with an array ID value then only that job array element will be canceled.

• If the job ID of a job array is specified without an array ID value then all job array elements will be canceled.

• A job or job step can only be canceled by the owner of that job or user root.
SLURM Commands

squeue
See job information with `squeue`

- `squeue` is used to view job and job step information for jobs

### Options

<table>
<thead>
<tr>
<th>Options</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>--user=vunetid</td>
<td>Show jobs only for this user.</td>
</tr>
<tr>
<td>--account=group</td>
<td>Show jobs only for this group</td>
</tr>
<tr>
<td>--states=[running or pending]</td>
<td>Show only running or pending jobs</td>
</tr>
<tr>
<td>--format=&quot;%.10i %N&quot;</td>
<td>Show only jobid's and nodes</td>
</tr>
<tr>
<td>--long</td>
<td>Provide more job information</td>
</tr>
<tr>
<td>--start</td>
<td>Sometimes gives approximate start time</td>
</tr>
</tbody>
</table>
SLURM Commands

sacct
Account information with sacct

- sacct is used to view accounting data for jobs.

<table>
<thead>
<tr>
<th>Options</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>--starttime 05.15.15</td>
<td>Shows info for all jobs starting on or after this date.</td>
</tr>
<tr>
<td>--j jobid</td>
<td>Shows info for this job.</td>
</tr>
<tr>
<td>--allusers</td>
<td>Shows info for all users. The default is current user.</td>
</tr>
<tr>
<td>--accounts=group</td>
<td>Shows info for this group.</td>
</tr>
<tr>
<td>--format=&quot;JobID,user,ncpus&quot;</td>
<td>Shows just jobid, user, and number of cpus.</td>
</tr>
<tr>
<td>--format has many additional specifications</td>
<td></td>
</tr>
</tbody>
</table>
SLURM Commands

rtracejob
• Quickly show important information about a job.

<table>
<thead>
<tr>
<th>User: autocms</th>
<th>JobID: 3115504</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account</td>
<td>cms_stage2</td>
</tr>
<tr>
<td>Job Name</td>
<td>skim_test</td>
</tr>
<tr>
<td>State</td>
<td>Completed</td>
</tr>
<tr>
<td>Exit Code</td>
<td>0:0</td>
</tr>
<tr>
<td>Wall Time</td>
<td>12:00:00</td>
</tr>
<tr>
<td>Requested Memory</td>
<td>2Gc</td>
</tr>
<tr>
<td>Memory Used</td>
<td>505864K</td>
</tr>
<tr>
<td>CPUs Requested</td>
<td>1</td>
</tr>
<tr>
<td>CPUs Used</td>
<td>1</td>
</tr>
<tr>
<td>Nodes</td>
<td>1</td>
</tr>
<tr>
<td>Node List</td>
<td>vmp736</td>
</tr>
<tr>
<td>Wait Time</td>
<td>1.5 minutes</td>
</tr>
<tr>
<td>Run Time</td>
<td>30.4 minutes</td>
</tr>
<tr>
<td>Submit Time</td>
<td>Tue Jun 30 14:06:03 2015</td>
</tr>
<tr>
<td>Start Time</td>
<td>Tue Jun 30 14:06:03 2015</td>
</tr>
<tr>
<td>End Time</td>
<td>Tue Jun 30 14:09:25 2015</td>
</tr>
<tr>
<td>Today's Date</td>
<td>Thu Jul  2 10:27:42 2015</td>
</tr>
</tbody>
</table>
Advanced Computing Center for Research and Education

SLURM Commands

scontrol

www.accre.vanderbilt.edu

Follow us on twitter: @ACCREVandy

GitHub repos: https://github.com/accre
### Interact with jobs using scontrol

<table>
<thead>
<tr>
<th>Command</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>scontrol show job [jobid]</td>
<td>Show job information yet another way.</td>
</tr>
<tr>
<td>scontrol hold [jobid]</td>
<td>Prevent a job from running.</td>
</tr>
<tr>
<td>scontrol release [jobid]</td>
<td>Release hold so job may be scheduled to run.</td>
</tr>
<tr>
<td>scontrol show nodes</td>
<td>Show information about compute nodes.</td>
</tr>
<tr>
<td>scontrol update dependency=jobid</td>
<td>Allow job to run after “jobid” completes.</td>
</tr>
</tbody>
</table>

Many `scontrol` commands can only be used by ACCRE staff. Adding time to a job is a good example of this.
SLURM Commands

salloc

Advanced Computing Center for Research and Education

www.accre.vanderbilt.edu

Follow us on twitter: @ACCREVandy

GitHub repos: https://github.com/accre
Run an interactive job with `salloc`

- Obtain a SLURM job allocation
- May be one or more nodes
- Often used for debugging.

`salloc --ntasks 1 --time=1:00:00`

This opens a shell on a node that a user may use for an hour.
Other useful commands

www.accre.vanderbilt.edu

Follow us on twitter: @ACCREVandy

GitHub repos: https://github.com/accre
ACCRE cluster commands

- **q3**
  - Shows the status of all jobs in the last hour arranged by users, groups and accounts.

- **qSummary**
  - Shows the number of user jobs arranged by groups.

- **showLimits**
  - Shows all the limits imposed on users and groups.

- **SlurmActive.pl**
  - Shows status of the cluster by node type.
Advanced Computing Center for Research and Education

Getting help from ACCRE

www.accre.vanderbilt.edu

Follow us on twitter: @ACCREVandy

GitHub repos: https://github.com/accre
Lots of help for you!

ACCRE website FAQ and Getting Started pages:
- http://www.accre.vanderbilt.edu/support

ACCRE Help Desk:
- http://www.accre.vanderbilt.edu/support/contact/submit_RT.php
- Tickets default to non-rush queue, which is for tickets about an issue which only impacts you and / or which can wait until the next business day.
- Rush tickets are for issues which would impact the cluster as a whole.
- Rush queue pages ACCRE personnel (we normally only have staff on site during normal business hours), so please think before opening up a rush ticket at 3 AM!
- If you open up a rush ticket - day or night - please plan on being available to work with ACCRE personnel to resolve the issue.
- Once the issue (rush or non-rush) is resolved we would appreciate it if you would let us know.
Lots of help for you!

SchedMD website for complete SLURM documentation.
  • http://slurm.schedmd.com/

• Clone our SLURM repo for job many script examples.
  • https://github.com/accre/SLURM

• Join our news groups for information about using R, matlab, python and ruby on the cluster. For details visit:
  • http://lists.accre.vanderbilt.edu/mailman/listinfo/

• For a handy list of SLUM commands check out:
  • http://slurm.schedmd.com/pdfs/summary.pdf