

DRAFT | Peanut Job Submission Cluster

We are currently **alpha** testing and gauging user interest in a cluster of machines that allows for the submission of long running compute jobs. Think of these machines as a dumping ground for discrete computing tasks that might have been rude or disruptive to execute on the main (shared) shell servers (i.e., linux1, linux2, linux3).

For job submission we will be using a piece of software called [SLURM](#). Simply put, SLURM is a queue management system and stands for **S**imple **L**inux **U**tility for **R**esource **M**anagement; it was developed at the Lawrence Livermore National Lab. It currently supports some of the largest compute clusters in the world. The best description of SLURM can be found on its homepage:

"Slurm is an open-source workload manager designed for Linux clusters of all sizes. It provides three key functions. First it allocates exclusive and/or non-exclusive access to resources (computer nodes) to users for some duration of time so they can perform work. Second, it provides a framework for starting, executing, and monitoring work (typically a parallel job) on a set of allocated nodes. Finally, it arbitrates contention for resources by managing a queue of pending work."¹⁾

SLURM is similar to most other queue systems in that you write a batch script, then submit it to the queue manager. The queue manager schedules your job to run on the queue (or partition in SLURM parlance) that you designate. Below is an outline of how to submit jobs to SLURM, how SLURM decides when to schedule your job, and how to monitor progress.

Where to begin

SLURM is a set of command line utilities that can be accessed via the command line from **most** any computer science system you can login to. Using our main shell servers (linux.cs.uchicago.edu) is expected to be our most common use case, so you should start there.

```
ssh user@linux.cs.uchicago.edu
```

Documentation

The [SLURM website](#) should be your primary source for documentation. If you Google SLURM questions, you'll often see the outdated Lawrence Livermore pages.

A great way to get details on SLURM commands are the manuals that are already on the cluster. For example, if you type the following command:

```
man sbatch
```

you will get the manual page for the sbatch command.

Resources

- [Common SLURM commands](#)
- [Official SLURM website](#)
- [Official SLURM documentation](#)
- [SLURM tutorial videos](#)
- [LLNL quick start user guide](#)

Infrastructure

Hardware

Our cluster contains nodes with the following specs:

- 16 Cores (2x 8core 3.1GHz Processors), 16 threads
- 64gb RAM
- 2x 500GB SATA 7200RPM in RAID1

To better manage the cluster we have virtualized the job submission nodes and give them all resources of the hardware. So, the actual resources you can consume on any one node is:

- 14 Cores, 14 threads
- 62GB RAM

Storage

There is slow scratch space mounted to `/scratch`. It is a ZFS pool consisting of 10x 2TB 7200RPM SAS drives connected via a LSI 9211-8i and is made up of 5 mirrored VDEVs, which is similar to a RAID10. The servers uplink is 1G ethernet.

- Files older than 90 days will be deleted automatically.
- Scratch space is shared by all users.

Access

Scratch space is only mounted on nodes associated with the cluster. If you want to be able to transfer files to the scratch space you will want to run an [interactive shell](#). Now you will be able to use standard tools such as `scp` or `rsync` to transfer files.

1. You should only do a file transfer via the debug partition: `srun -p debug --pty --mem 500 /bin/bash`
2. Now you can create a directory of your own: `mkdir /scratch/$USER` You should store any files you create in this directory.

Example

Request interactive shell

```
user@csilcomputer:~$ srun --pty --mem 500 /bin/bash
```

Change into my scratch directory:

```
user@research2:~$ cd /scratch/user/
```

Get the files I need:

```
user@research2:/scratch/user$ scp user@csilcomputer:~/foo .
foo                                100% 103KB 102.7KB/s   00:00
```

Check that the file now exists:

```
user@research2:/scratch/user$ ls -l foo
-rw----- 1 user user 105121 Dec 29 2015 foo
```

I can now exit my interactive shell.

Performance is slow

This is expected. The maximum speed this server will ever be able to achieve is 1Gb/s because of its single 1G ethernet uplink. If this cluster gains in popularity we plan on upgrading the network and storage server.

Utilization Dashboard

Sometimes it is useful to see how much of the cluster is utilized. You can do that via the following URL: <http://peanut.cs.uchicago.edu>

Partitions / Queues

To find out what partitions we offer, checkout the [sinfo](#) command.

As of December, 2015 we have will have at least 2 partitions in our cluster; 'debug' and 'general'.

Partition Name	Description
debug	The partition your job will be submitted to if none is specified. The purpose of this partition is to make sure your code is running as it should before submitting a long running job to the general queue.

Partition Name	Description
general	All jobs that have been thoroughly tested can be submitted here. This partition will have access to more nodes and will process most of the jobs. If you need to use the <code>--exclusive</code> flag it should be done here.
gpu	Contains servers with graphics cards. As of May 2016 there is only one node containing a Tesla M2090. You will be forced to use this server exclusively for now. Please keep your time in interactive mode to a minimum.

Job Submission

Jobs submitted to the cluster are run from the command line. Almost anything that you can run via the command line on any of our machines in our labs can be run on our job submission server agents.

The job submission servers run Ubuntu 14.04 with the same software as you will find on our lab computers, but without the X environment.

You can submit jobs from the departmental computers that you have access to. You will not be able to access the job server agent directly.

Command Summary

[Cheat Sheet](#)

	SLURM	Example
Submit a batch serial job	<code>sbatch</code>	<code>sbatch runscript.sh</code>
Run a script interactively	<code>srun</code>	<code>srun -pty -p interact -t 10 -mem 1000 /bin/bash /bin/hostname</code>
Kill a job	<code>scancel</code>	<code>scancel 4585</code>
View status of queues	<code>squeue</code>	<code>squeue -u cnetid</code>
Check current job by id	<code>sacct</code>	<code>sacct -j 999999</code>

Usage

Below are some common examples. You should consult the [documentation](#) of SLURM if you need further assistance.

Exclusive access to a node

You will need to add the `--exclusive` options to your script or command line options. This option will ensure that when your job runs it is the only job running on that particular node.

SBATCH

The sbatch command is used for submitting jobs to the cluster. sbatch accepts a number of options either from the command line, or (more typically) from a batch script. An example of a SLURM batch script is shown below:

Sample script

Make sure you create a directory in which to deposit the STDIN, STDOUT, STDERR files.

```
mkdir -p $HOME/slurm/slurm_out
```

```
#!/bin/bash
#
#SBATCH --mail-user=cnetid@cs.uchicago.edu
#SBATCH --mail-type=ALL
#SBATCH --output=/home/cnetid/slurm/slurm_out/%j.%N.stdout
#SBATCH --error=/home/cnetid/slurm/slurm_out/%j.%N.stderr
#SBATCH --workdir=/home/cnetid/slurm
#SBATCH --partition=debug
#SBATCH --job-name=check_hostname_of_node
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --time=15:00

hostname
```

If any of the above options are unclear as to what they do please check the man page for sbatch

```
man sbatch
```

Make sure to replace all instances of the word cnetid with your CNETID.

Submitting job script

Using the above example you will want to place your tested code into a file. 'hostname.job' is the file name in this example.

```
sbatch hostname.job
```

You can then check the status via squeue or see the output in the output directory '\$HOME/slurm/slurm_out'.

SRUN

Used to submit a job to the cluster that doesn't necessarily need a script.

```
user@host:~$ srun -n2 hostname
research2
research2
```

`srun` will remain in the foreground until the job has finished.

```
user@host:~$ srun -n1 sleep 400
```

squeue

This command will show jobs in the queue.

```
user@host:~$ squeue
JOBID PARTITION    NAME    USER ST      TIME  NODES NODELIST(REASON)
   29      debug    sleep    user  R       0:11      1 research2
```

scancel

Cancel one of your own jobs. Please read the `scancel` manual page (`man scancel`) as there are many ways of canceling your jobs if they are of any complexity.

```
scancel 29
```

sinfo

View information about SLURM nodes and partitions.

The following code block shows the what happens when you run the `sinfo` command. You get a list of 'partitions' on which you can run your code. Each partition is comprised of certain types of nodes. In the case below the default (denoted by a `*`) is 'debug'. The job time limit is short and is meant only to debug your code. The other partitions will usually have a particular purpose in mind. 'hardware', for example, is to be used if you require direct access to the hardware instead of the KVM layer between the hardware and the OS.

```
user@host:~$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
debug*    up      30:00      1    idle research2
general   up    1-00:00:00      6    idle research[3-8]
```

Monitoring Jobs

`squeue` and `sacct` are two different commands that allow you to monitor job activity in SLURM. `squeue` is the primary and most accurate monitoring tool since it queries the SLURM controller directly. `sacct` gives you similar information for running jobs, and can also report on previously

finished jobs, but because it accesses the SLURM database, there are some circumstances when the information is not in sync with `squeue`.

Running `squeue` without arguments will list all currently running jobs. It is more common, though to list jobs for a particular user (like yourself) using the `-u` option...

```
squeue -u cnetid
```

or for a particular job id.

```
squeue -j 7894
```

Interactive Jobs

Though batch submission is the best way to take full advantage of the compute power in the job submission cluster, foreground, interactive jobs can also be run.

An interactive job differs from a batch job in two important aspects:

1. The partition to be used is the `interact` partition
2. Jobs should be initiated with the `srun` command instead of `sbatch`.

This command:

```
srun -p general --pty --cpus-per-task 1 --mem 500 -t 0-06:00 /bin/bash
```

will start a command line shell (`/bin/bash`) on the 'general' queue with 500 MB of RAM for 6 hours; 1 core on 1 node is assumed as these parameters (`-n 1 -N 1`) were left out. When the interactive session starts, you will notice that you are no longer on a login node, but rather one of the compute nodes dedicated to this queue. The `--pty` option allows the session to act like a standard terminal.

Job Scheduling

We use a [multifactor](#) method of job scheduling. Job priority is assigned by a combination of fair-share, partition priority, and length of time a job has been sitting in the queue. The priority of the queue is the highest factor in the job priority calculation. For certain queues this will cause jobs on lower priority queues which overlap with that queue to be requeued. The second most important factor is fair-share score. You can find a description of how SLURM calculates Fair-share [here](#). The third most important is how long you have been sitting in the queue. The longer your job sits in the queue the higher its priority grows. If everyone's priority is equal then FIFO is the scheduling method. If you want to see what your current priority is just do `sprio -j JOBID` which will show you the calculation it does to figure out your job priority. If you do `sshare -u USERNAME` you can see your current fair-share and usage.²⁾

We also have backfill turned on. This allows for jobs which are smaller to sneak in while a larger higher priority job is waiting for nodes to free up. If your job can run in the amount of time it takes for

the other job to get all the nodes it needs, SLURM will schedule you to run during that period. **This means knowing how long your code will run for is very important and must be declared if you wish to leverage this feature. Otherwise the scheduler will just assume you will use the maximum allowed time for the partition when you run.**³⁾

Common Issues

Error	What does it mean?
JOB <jobid> CANCELLED AT <time> DUE TO TIME LIMIT	You did not specify enough time for your job to run. The -t flag will allow you to set the time limit.
Job <jobid> exceeded <mem> memory limit, being killed	Your job is attempting to use more memory that you have requested for it. Either increase the amount of memory you have requested or reduce the amount of memory usage your application is trying to use.
JOB <jobid> CANCELLED AT <time> DUE TO NODE FAILURE	There can be many reasons for this message, but most often it means that the node your job was set to run on can no longer be contacted by the the SLURM controller.
error: Unable to allocate resources: More processors requested than permitted	It usually has nothing to do with priviledges you may or may not have. Rather, it usually means that you have allocated more processors than one compute node actually has.

Using the GPU

Example

This sbatch script will get device information from the installed Tesla gpu.

```
#!/bin/bash
#
#SBATCH --mail-user=cnetid@cs.uchicago.edu
#SBATCH --mail-type=ALL
#SBATCH --output=/home/cnetid/slurm/slurm_out/%j.%N.stdout
#SBATCH --error=/home/cnetid/slurm/slurm_out/%j.%N.stderr
#SBATCH --workdir=/home/cnetid/slurm
#SBATCH --partition=gpu
#SBATCH --job-name=get_tesla_info

cat << EOF > /tmp/getinfo.cu
#include <stdio.h>

int main() {
    int nDevices;

    cudaGetDeviceCount(&nDevices);
    for (int i = 0; i < nDevices; i++) {
        cudaDeviceProp prop;
```



```
    cudaGetDeviceProperties(&prop, i);
    printf("Device Number: %d\n", i);
    printf("  Device name: %s\n", prop.name);
    printf("  Memory Clock Rate (KHz): %d\n",
           prop.memoryClockRate);
    printf("  Memory Bus Width (bits): %d\n",
           prop.memoryBusWidth);
    printf("  Peak Memory Bandwidth (GB/s): %f\n\n",
           2.0*prop.memoryClockRate*(prop.memoryBusWidth/8)/1.0e6);
  }
}
EOF

/usr/local/cuda/bin/nvcc /tmp/getinfo.cu -o /tmp/a.out
/tmp/a.out
rm /tmp/a.out
rm /tmp/getinfo.cu
```

Output

STDOUT will look something like this:

```
cnetid@linux1:~$ cat $HOME/slurm/slurm_out/12567.gpu1.stdout
Device Number: 0
  Device name: Tesla M2090
  Memory Clock Rate (KHz): 1848000
  Memory Bus Width (bits): 384
  Peak Memory Bandwidth (GB/s): 177.408000
```

STDERR should be blank.

More

If you feel this documentation is lacking in some way please let techstaff know. Email techstaff@cs.uchicago.edu, call (773-702-1031), or stop by our office (Ryerson 154).

1)

<http://slurm.schedmd.com/>

2) 3)

<https://rc.fas.harvard.edu/resources/running-jobs>

From:
<https://howto.cs.uchicago.edu/> - **How do I?**

Permanent link:
<https://howto.cs.uchicago.edu/techstaff:slurm?rev=1462824518>

Last update: **2016/05/09 15:08**



