



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

CRUK CI HPC cluster introduction (II of III)

Using the scheduler for job submission



UNIVERSITY OF
CAMBRIDGE

Overview

SLURM (Simple Lightweight Unix Resource Manager)

In their own words:

'Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters. Slurm requires no kernel modifications for its operation and is relatively self-contained. As a cluster workload manager, Slurm has three key functions. First, it allocates exclusive and/or non-exclusive access to resources (compute nodes) to users for some duration of time so they can perform work. Second, it provides a framework for starting, executing, and monitoring work (normally a parallel job) on the set of allocated nodes. Finally, it arbitrates contention for resources by managing a queue of pending work.'

SLURM will allow you to:

- Submit jobs to the cluster.
- Specify which queue/account to submit your jobs to.
- Request memory resources for your jobs.
- Set memory limits for your jobs.
- Set time limits for your jobs
- Check the status of the jobs you have submitted.
- Check the status of the hosts within the cluster.
- Kill jobs that you have submitted to the cluster.

The most useful SLURM commands

These are:

- **srun**
- **sbatch**
- **sacct**
- **squeue**
- **scancel**

Type command followed by -h for usage details.

Usage: srun [OPTIONS...] executable [args...]

Parallel run options:

-A, --account=name charge job to specified account
--acctg-freq=<datatype>=<interval> accounting and
profiling sampling intervals. Supported datatypes:
task=<interval> energy=<interval>
network=<interval> filesystem=<interval>

...

Help options:

-h, --help show this help message
--usage display brief usage message

Other options:

-V, --version output version information and exit

A Job

A command or series of commands submitted to the cluster with associated resource requirements and limits.

Status of jobs running on the cluster can be seen with the command

clust1-headnode \$ squeue

```
[adm-ct@clust1-headnode ~]$ squeue
JOBID PARTITION NAME USER ST TIME NODES NODELIST(REASON)
476136 general detect_a morris01 R 19:49:27 1 clust1-node-22
476137 general detect_a morris01 R 19:49:27 1 clust1-node-23
476138 general detect_a morris01 R 19:49:27 1 clust1-node-24
476139 general detect_a morris01 R 19:49:27 1 clust1-node-25
476140 general detect_a morris01 R 19:49:27 1 clust1-node-26
480480 general bash sawle01 R 44:34 1 clust1-node-29
480482 general bash lukk01 R 21:45 1 clust1-node-1
177846 general HLLVBBX solexa R 25-21:50:29 1 clust1-node-10
480479 general bash sammut01 R 1:32:53 1 clust1-node-16
480478 general bash sammut01 R 1:33:43 1 clust1-node-3
480476 general bash sammut01 R 1:40:56 1 clust1-node-7
480483 general wrap portel01 R 18:41 1 clust1-node-28
475686 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-17
475687 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-19
475688 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-20
475689 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-21
475690 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-22
475691 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-23
475692 general bKO(CG_1 portel01 R 20:30:18 1 clust1-node-24
475693 general bKO(CG_2 portel01 R 20:30:18 1 clust1-node-25
475694 general bKO(CG_2 portel01 R 20:30:18 1 clust1-node-26
475673 general bKO(CG_0 portel01 R 20:30:21 1 clust1-node-19
475674 general bKO(CG_1 portel01 R 20:30:21 1 clust1-node-30
475675 general bKO(CG_2 portel01 R 20:30:21 1 clust1-node-31
475676 general bKO(CG_3 portel01 R 20:30:21 1 clust1-node-2
475677 general bKO(CG_4 portel01 R 20:30:21 1 clust1-node-3
475678 general bKO(CG_5 portel01 R 20:30:21 1 clust1-node-4
475679 general bKO(CG_6 portel01 R 20:30:21 1 clust1-node-5
475680 general bKO(CG_7 portel01 R 20:30:21 1 clust1-node-11
475681 general bKO(CG_8 portel01 R 20:30:21 1 clust1-node-12
475682 general bKO(CG_9 portel01 R 20:30:21 1 clust1-node-13
475683 general bKO(CG_1 portel01 R 20:30:21 1 clust1-node-14
475684 general bKO(CG_1 portel01 R 20:30:21 1 clust1-node-15
475685 general bKO(CG_1 portel01 R 20:30:21 1 clust1-node-16
475587 general bKO(CG_0 portel01 R 20:30:52 1 clust1-node-6
475588 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-7
475589 general bKO(CG_2 portel01 R 20:30:52 1 clust1-node-8
475590 general bKO(CG_3 portel01 R 20:30:52 1 clust1-node-11
475591 general bKO(CG_4 portel01 R 20:30:52 1 clust1-node-12
475592 general bKO(CG_5 portel01 R 20:30:52 1 clust1-node-14
475593 general bKO(CG_6 portel01 R 20:30:52 1 clust1-node-16
475594 general bKO(CG_7 portel01 R 20:30:52 1 clust1-node-18
475595 general bKO(CG_8 portel01 R 20:30:52 1 clust1-node-20
475596 general bKO(CG_9 portel01 R 20:30:52 1 clust1-node-21
475597 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-22
475598 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-23
475599 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-24
475600 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-27
475601 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-32
475602 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-33
475603 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-4
475604 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-5
475605 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-6
475606 general bKO(CG_1 portel01 R 20:30:52 1 clust1-node-7
475607 general bKO(CG_2 portel01 R 20:30:52 1 clust1-node-8
475608 general bKO(CG_2 portel01 R 20:30:52 1 clust1-node-9
475609 general bKO(CG_2 portel01 R 20:30:52 1 clust1-node-10
[adm-ct@clust1-headnode ~]$
```

A Queue:

A queue for job submissions associated with specified users and cluster hosts, and providing specified default resources.

SLURM calls queues ‘partitions.’

We have a single general ‘partition’, with few restrictions (currently).

Checking the status of hosts within the cluster

SLURM

Clust1-headnode > sinfo

```
[obrien04@clust1-headnode ~]$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
general*    up    infinite     15    mix  clust1-node-[1-7,9,12-16,18,21]
general*    up    infinite     18    idle clust1-node-[8,10-11,17,19-20,22-33]
```

Killing jobs with scancel

```
clust1-headnode $ scancel <job-id>
```

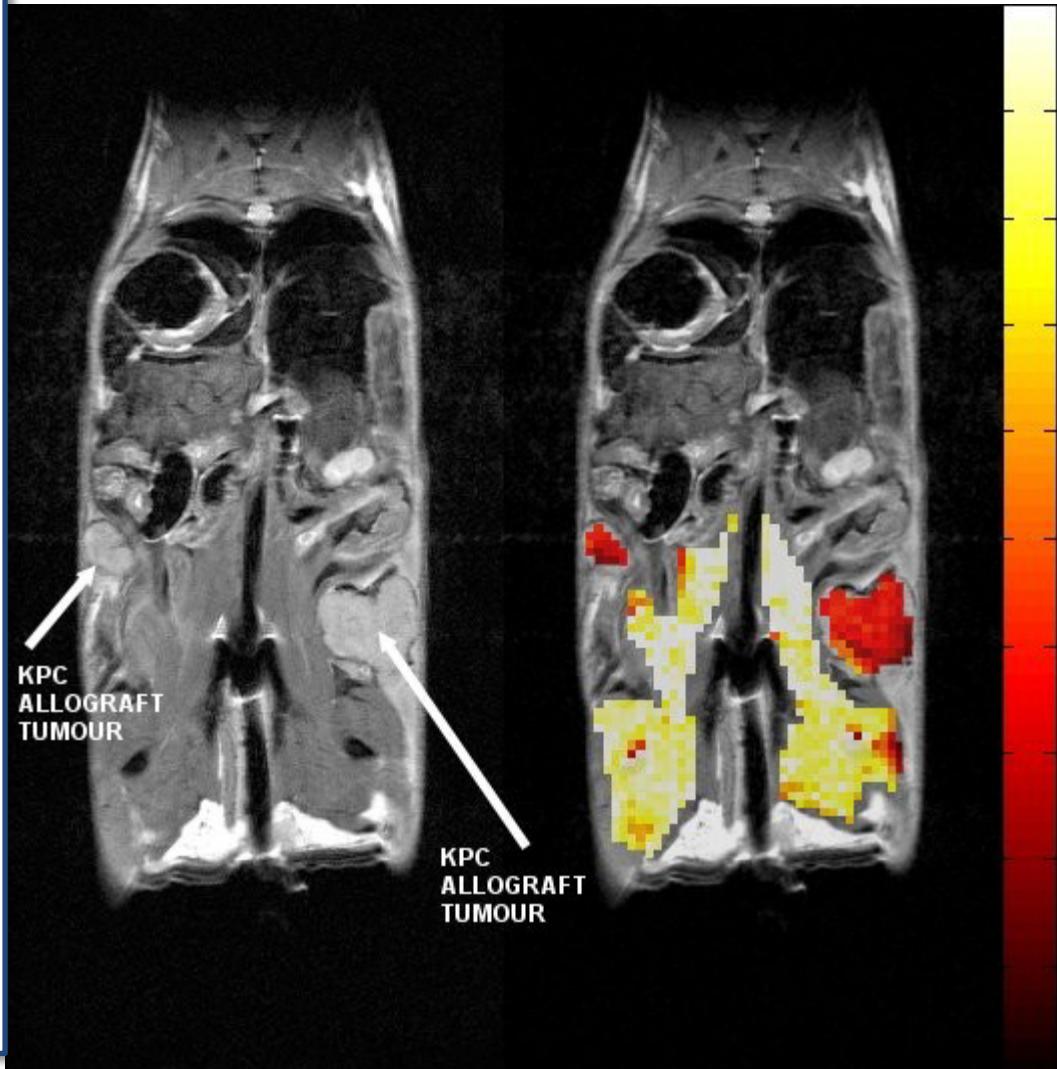
Simple Parallel Computing

Most HPC at CRI consists of breaking up your dataset into chunks and firing off a separate job for each chunk.

'Magnetisation Transfer Map Example'

Here small groups of pixels from T2 weighted image slices are processed to calculate the elements of the MT map.'

(Example courtesy of Dominick McIntyre,
Griffiths Group)



Job array example

Adapted from an example by Stewart MacArthur, Bioinformatics Core Facility

'Here is a self contained example of the basics of using job arrays. The main benefit in this case is not speed but the ability to control the number of jobs running at any one time, using the %50 notation in the sbatch... I find job arrays particularly useful for running lots of small jobs, as there is only a single job submission there is little SLURM overhead, compared to submitting 1000 separate jobs, which takes some time. Also being able to control the number of running jobs stops you swamping your queue with jobs and leaves space for others to get jobs running.'

```
### Generate a random big file that we want to sort, 10 Million lines
perl -e 'for (1..1E7){printf("%.0f\n",rand()*1E7)};' > bigFile

### Split the file up into chunks with 10,000 lines in each chunk
split -a 3 -d -l 10000 bigFile split

### rename the files on a 1-1000 scheme not 0-999
for f in split*;do mv ${f} $(echo ${f} |perl -ne 'm/split(0*) (\d+)/g;print "Split",${2+1},"\n";');done

### submit a job array, allowing 50 jobs to be run at anyone time
sbatch --time=1:0 --array=1-1000%50 -N1 ./sort-script.sh

### merge the sorted files together once all the jobs are finished
sort -n -m *.sorted > bigFile.sorted

### Delete the temp files
rm -f Split*
```

```
[thomso04@clust1-headnode-1 /home/thomso04]# cat sort-script.sh
#!/bin/bash
/usr/bin/sort -n /scratcha/computing/Split$SLURM_ARRAY_TASK_ID > /scratcha/computing/Split$SLURM_ARRAY_TASK_ID.sorted
```

Job dependency

SLURM allows many different ways to express dependencies, using the `--dependency` switch

Some SLURM dependencies:

`after:job_id[:jobid...]`

This job can begin execution after the specified jobs have begun execution.

`afterany:job_id[:jobid...]`

This job can begin execution after the specified jobs have terminated.

`afterok:job_id[:jobid...]`

This job can begin execution after the specified jobs have successfully executed (ran to completion with an exit code of zero).

```
[clust1-headnode ~] $ sbatch job1.sh  
11254323
```

```
[clust1-headnode ~] $ sbatch --dependency=afterok:11254323 job2.sh
```



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Practical session II



UNIVERSITY OF
CAMBRIDGE