SLURM

What is Slurm?

- Slurm is an open-source job scheduler and cluster maintainer.
- Built collaboratively by Lawrence Livermore National Laboratory, SchedMD, Linux NetworX, Hewlett-Packard, and Groupe Bull
- Inspired by Quadratics RMS
- Approximately 60% of the TOP500 supercomputers use Slurm as their cluster manager
- Uses a best-fit algorithm
- Fair-share scheduling
- Highly scalable
- Supports job arrays and job dependencies

Common Commands

- sbatch
- squeue
- scontrol
- scancel
- sinfo
- sacct
- srun

sbatch vs qsub

sbatch	qsub
#SBATCH	#$
-t 00:15:00	-l h_rt=00:15:00
--mem=2GB	-l mem_free=2G
-p all.q	-q all.q
-N 4 --ntasks-per-node 4	-pe fixed4 16
-A myGrp	-P myPrj
-D ./	-cwd

https://slurm.schedmd.com/sbatch.html

squeue

- Job status
- squeue –j xxxxxx
- squeue –u *juser*

```
[cwf25@login005 python]$ sbatch testJob.sh
Submitted batch job 9816307
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816307        RM testJob.    cwf25 PD       0:00      1 (None)
           9816306        RM testJob.    cwf25  R       0:10      1 r741
[cwf25@login005 python]$ squeue -j 9816306
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816306        RM testJob.    cwf25  R       0:19      1 r741
[cwf25@login005 python]$
```

https://slurm.schedmd.com/squeue.html

States

- CA - cancelled
- CG – completing
- CD – completed
- F – job failed
- PD – pending
- R – running
- RQ – requeued
- S – suspended
- TO – time out

scontrol

- scontrol used to modify or view Slurm configuration
- scontrol show job xxxxxx

```
[cwf25@login005 python]$ scontrol show job 9816306
JobId=9816306 JobName=testJob.sh
   UserId=cwf25(75249) GroupId=tr5fpqp(23162) MCS_label=N/A
   Priority=3888 Nice=0 Account=tr5fpqp QOS=rm
   JobState=COMPLETED Reason=None Dependency=(null)
   Requeue=0 Restarts=0 BatchFlag=1 Reboot=0 ExitCode=0:0
   RunTime=00:02:50 TimeLimit=01:00:00 TimeMin=N/A
   SubmitTime=2020-06-23T16:00:46 EligibleTime=2020-06-23T16:00:46
   AccrueTime=2020-06-23T16:00:46
   StartTime=2020-06-23T16:01:18 EndTime=2020-06-23T16:04:08 Deadline=N/A
   PreemptTime=None SuspendTime=None SecsPreSuspend=0
   LastSchedEval=2020-06-23T16:01:18
   Partition=RM AllocNode:Sid=br005:29227
   ReqNodeList=(null) ExcNodeList=(null)
   NodeList=r741
   BatchHost=r741
   NumNodes=1 NumCPUs=28 NumTasks=4 CPUs/Task=1 ReqB:S:C:T=0:0:*:*
   TRES=cpu=28,mem=2G,node=1,billing/gpu=28
   Socks/Node=* NtasksPerN:B:S:C=4:0:*:* CoreSpec=*
   MinCPUsNode=4 MinMemoryNode=2G MinTmpDiskNode=0
   Features=(null) DelayBoot=00:00:00
   OverSubscribe=NO Contiguous=0 Licenses=(null) Network=(null)
   Command=/pylon5/tr5fpqp/cwf25/tests/python/testJob.sh
   WorkDir=/pylon5/tr5fpqp/cwf25/tests/python
   StdErr=/pylon5/tr5fpqp/cwf25/tests/python/slurm-9816306.out
   StdIn=/dev/null
   StdOut=/pylon5/tr5fpqp/cwf25/tests/python/slurm-9816306.out
   Power=
```

https://slurm.schedmd.com/scontrol.html

scontrol

- scontrol hold xxxxxx
- scontrol release xxxxxx

```
[cwf25@login005 python]$ sbatch testJob.sh
Submitted batch job 9816364
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816364        RM testJob.    cwf25 PD       0:00      1 (None)
[cwf25@login005 python]$ scontrol hold 9816364
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816364        RM testJob.    cwf25 PD       0:00      1 (JobHeldUser)
[cwf25@login005 python]$ scontrol release 9816364
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816364        RM testJob.    cwf25 PD       0:00      1 (None)
[cwf25@login005 python]$
```

https://slurm.schedmd.com/scontrol.html

scontrol

- scontrol show nodes
  - scontrol show node <*node*>
- scontrol show partition <*partition*>

```
[cwf25@login005 python]$ scontrol show nodes
NodeName=dgpu001 Arch=x86_64 CoresPerSocket=16
   CPUAlloc=16 CPUTot=32 CPULoad=0.26
   AvailableFeatures=EGRESS,PERF,E5-2683,E5-2683v4,p100,DBMI
   ActiveFeatures=EGRESS,PERF,E5-2683,E5-2683v4,p100,DBMI
   Gres=gpu:p100:2
   NodeAddr=dgpu001 NodeHostName=dgpu001 Version=18.08
   OS=Linux 3.10.0-957.27.2.el7.x86_64 #1 SMP Mon Jul 29 17:46:05 UTC 2019
   RealMemory=128000 AllocMem=61600 FreeMem=112022 Sockets=2 Boards=1
   State=MIXED ThreadsPerCore=1 TmpDisk=0 Weight=1000 Owner=N/A MCS_label=N/A
   Partitions=DBMI-GPU
   BootTime=2020-05-28T12:33:36 SlurmdStartTime=2020-05-28T12:34:59
   CfgTRES=cpu=32,mem=125G,billing/gpu=32,gres/gpu=2,gres/gpu:p100=2
   AllocTRES=cpu=16,mem=61600M,gres/gpu=1,gres/gpu:p100=1
   CapWatts=n/a
   CurrentWatts=0 LowestJoules=0 ConsumedJoules=0
   ExtSensorsJoules=n/s ExtSensorsWatts=0 ExtSensorsTemp=n/s
```

https://slurm.schedmd.com/scontrol.html

scancel

- Delete a job
- scancel xxxxxx

```
[cwf25@login005 python]$ sbatch testJob.sh
Submitted batch job 9816133
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           9816133        RM testJob.    cwf25 PD       0:00      1 (None)
[cwf25@login005 python]$ scancel 9816133
[cwf25@login005 python]$ squeue -u cwf25
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
[cwf25@login005 python]$
```

https://slurm.schedmd.com/scancel.html

sinfo

- Provides information about nodes and partitions managed by Slurm

```
[cwf25@login005 python]$ sinfo -s
PARTITION  AVAIL  TIMELIMIT   NODES(A/I/O/T)  NODELIST
RM*           up 3-00:00:00    426/311/2/739  r[006-744]
RM-shared     up 3-00:00:00     160/39/1/200  r[553-752]
RM-small      up    8:00:00          2/3/0/5  r[001-005]
GPU           up 2-00:00:00        44/0/0/44  gpu[001-044]
GPU-shared    up 2-00:00:00        44/0/0/44  gpu[001-044]
GPU-small     up    8:00:00          0/4/0/4  gpu[045-048]
GPU-AI        up 2-00:00:00         4/3/3/10  gpu[049-058]
LM            up 14-00:00:0        46/0/0/46  l[001-042],xl[001-004]
XLM           up 14-00:00:0          4/0/0/4  xl[001-004]
DBMI          up 2-00:00:00          8/0/0/8  dr[001-008]
DBMI-GPU      up 2-00:00:00          2/2/0/4  dgpu[001-004]
```

NODES(A/I/O/T)
A – Available
I – Idle
O – Other
T – Total

https://slurm.schedmd.com/sinfo.html

sacct

- Displays accounting information for Slurm
- Fields can be specified with --format
  - sacct –j xxxxxx --format=JobID,JobName,Account,NTasks,...
  - To see available fields type "sacct –e"

```
[cwf25@login005 python]$ sacct -j 9816364
       JobID    JobName  Partition    Account  AllocCPUS      State ExitCode
------------ ---------- ---------- ---------- ---------- ---------- --------
9816364      testJob.sh         RM    tr5fpqp         28  COMPLETED      0:0
9816364.bat+      batch               tr5fpqp         28  COMPLETED      0:0
9816364.ext+     extern               tr5fpqp         28  COMPLETED      0:0
[cwf25@login005 python]$
```

https://slurm.schedmd.com/sacct.html

srun

- Start an interactive job
  - srun <args> --pty bash
- Also used to submit jobs
- A sbatch may have multiple srun commands within it

```
[cwf25@login005 python]$ srun -N 1 --ntasks-per-node=4 -t 00:05:00 --pty bash
srun: job 9816772 queued and waiting for resources
srun: job 9816772 has been allocated resources
[cwf25@r741 python]$
```

https://slurm.schedmd.com/srun.html

Array Jobs

- --array=n[-m[:s]]
  - n – start ID
  - m – end ID
  - s – step
- You must know how many tasks before you submit
- --requeue will restart jobs that end unexpectedly

Env. Variable	Meaning
SLURM_ARRAY_JOB_ID	First job ID of the array
SLURM_ARRAY_TASK_ID	Job array index value
SLURM_ARRAY_TASK_MAX	Highest job array index value
SLURM_ARRAY_TASK_MIN	Lowest job array index value
SLURM_ARRAY_TASK_COUNT	Number of tasks in the job array

https://slurm.schedmd.com/job_array.html

Dependencies

- --dependency=<AFTER>:<JOB_ID>[:<JOB_ID>…]

- The <AFTER> field tells Slurm what to look for as a signal to start the job
  - after – Can begin after specified job starts running
  - afterok – Can begin after successful ending of specified job
  - afternotok – Can begin after specified job fails
  - afterany – Can begin after the specifed job ends, regardless of status

Let's Try It!

XSEDE