

SLURM: Workload manager for HPC batch cluster

S. Ravichandran PhD* and Dennis Foley⁺

*ABCS, ⁺IT Operations

LBR, FNLCR

Latest copy of the presentation can be obtained from

<https://github.com/FNLCR-Bioinformatics/ProgrammersCorner-Parallelization>

Agenda

- To introduce the SLURM scheduler/resource manager batch system replacing the current PBS job scheduler
 - Why the change?
- Key SLURM commands
 - Compare with PBS
- Couple of examples using SLURM

What is SLURM?

- **Simple Linux Unified Resource Manager**
 - Written in C (half a million-lines of code)
- Linux utility for Resource management for job scheduling and resource management
- Open source utility widely used in other computing centers
 - Research centers, across the globe

Why Enterprise Information Technology is migrating to SLURM?

- Make the new batch system compatible with CIT's Beowulf computing cluster
- Linux distribution OS will change from Ubuntu to CentOS in the next six months
- The proprietary Moab/Torque scheduler/resource manager will be replaced by SLURM during the migration

New System aiming to bring a large performance boost

Information gleaned from Dr. Doug O'Neal's email

- Servers
 - HPE DL360 Gen 14 servers (no hyper-threading)
- System
 - 2 Intel Skylake processors
 - 18 physical cores each running at 2.7GHz.
 - 768GB memory, SSD system disk, and a 960GB SSD scratch disk.
- Few systems
 - 2 Nvidia GTX 1080Ti graphics cards
 - Support image processing and machine learning efforts
- Contact x 5115 for any questions on the new batch

Useful commands

- **sinfo**: View information & status of SLURM nodes and partitions
 - `pbsnodes -a`
- **sbatch/scancel**: Submit/cancel a batch (script) to SLURM
 - `qsub/qdel`
- **sacct**: To display accounting data for all jobs
- **squeue**: To view information about all jobs located in SLURM scheduling queue
 - `squeue -u username`

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

sinfo: information on nodes & partitions

```
bash-4.2$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
quick      up       4:00:00    7    idle  Node1
norm*      up    10-00:00:0    6    idle  Node2
unlimited   up      infinite    2    idle  Node3
```

- **PARTITIONS** are equivalent to the old PBS batch queues
 - quick partition contains 7 nodes (for now, each node = 16 cores)
 - normal partition contains 6 nodes
 - unlimited partition contains 2 nodes
- **TIMELIMIT Column contains the maximum time allowed for each PARTITION**
 - 4:00:00 means hr:min:sec

Examples

How to submit a job using AMBER as an example

PBS → SLURM

PBS	SLURM	What it means?
qsub <job-file>	sbatch <job-file>	Submit the script to queue
qsub -I	salloc <options>	Request interactive job
showstart	squeue -start	Display estimated start time
qstat <-u username>	squeue <-u username>	Check jobs for username
qstat <queue>	squeue -p <partition>	Display queue/partition information
qstat -f <job-id>	scontrol show job <job-id>	Display detailed job details (WorkDir, .e or .o output file dir location etc.)
qdel <job-id>	scancel <job-id>	Delete <job-id>

```
#PBS -S /bin/csh -N VINA-1 -r n -c n
#PBS -l pvmem=2gb
#PBS -l walltime=18:10:00
#PBS -l nodes=1:ppn=16
```

```
#!/usr/bin/tcsh
#SBATCH --export=NONE
#SBATCH --partition=norm
#SBATCH --job-name=VINA-1
#SBATCH --output=VINA-1.out
#SBATCH --nodes=1
#SBATCH --ntasks=16
#SBATCH --time=18:10:00
#SBATCH --mem-per-cpu=2gb
```

sinfo

Note: Node names are hidden
for security purposes

```
bash-4.2$ sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
quick      up       4:00:00    7   idle  Node1
norm*      up    10-00:00:0    6   idle  Node2
unlimited   up      infinite    2   idle  Node3
```

sbatch

scancel 532

```
[ravichandrans@[REDACTED] Amber_SLURM]$ sbatch amber16_tutorial_3.batch
Submitted batch job 532
```

squeue

squeue -u username

```
[ravichandrans@[REDACTED] Amber_SLURM]$ squeue
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST (REASON)
532	norm	AMBER16	username	R	0:50	1	[REDACTED]

sacct

```
[ravichandrans@[REDACTED] Amber_SLURM]$ sacct
```

JobID	JobName	Partition	Account	AllocCPUS	State	ExitCode
532	AMBER16 M+	norm		16	RUNNING	0:0

```
[ravichandrans@fr-s-hpc-a2-01 Vina_SLURM]$ cat Vina2018-1.pbs
```

```
#!/usr/bin/tcsh
```

```
#SBATCH --export=NONE
```

```
#SBATCH --partition=norm
```

```
#SBATCH --job-name=VINA-1
```

```
#SBATCH --output=VINA-1.out
```

```
#SBATCH --nodes=1
```

```
#SBATCH --ntasks=16
```

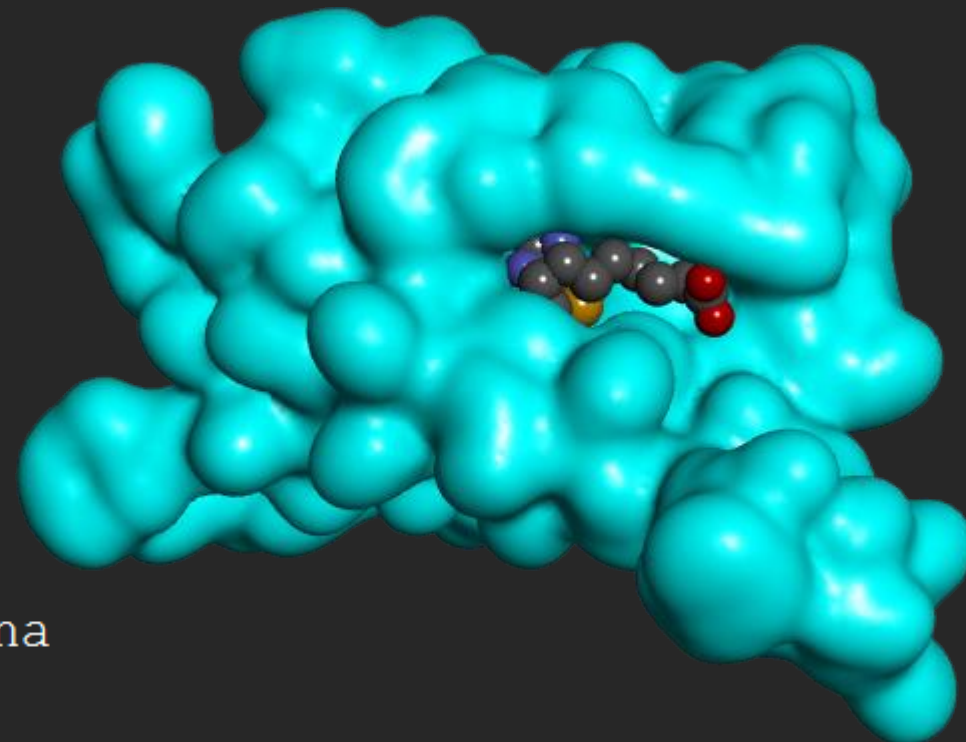
```
#SBATCH --time=18:10:00
```

```
#SBATCH --mem-per-cpu=2gb
```

```
setenv VINA autodock_vina_1_1_2_linux_x86/bin/vina
```

```
cd /users/priapp/$USER/SLURM/Vina_SLURM
```

```
$VINA/vina --config vina_conf.txt
```



```
[ravichandrans@ [REDACTED] Vina_SLURM]$ sbatch VINA.batch
```

```
Submitted batch job 534
```

```
[ravichandrans@ [REDACTED] Vina_SLURM]$ squeue
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST (REASON)
534	norm	VINA-1	username	R	0:08	1	[REDACTED]

Note: Node names are hidden
for security purposes

```
[ravichandrans@fr-s-hpc-a2-01 Vina_SLURM]$ tail VINA-1.out
```

2	-6.4	1.455	2.465
3	-6.3	1.316	2.442
4	-6.2	1.264	2.253
5	-5.8	1.192	1.791
6	-5.5	1.762	3.063
7	-4.6	2.734	6.611
8	-4.1	1.185	2.521
9	-3.8	1.485	2.275
10	-3.6	1.751	2.158

```
Writing output ... done.
```

Thanks to Dennis Foley for providing the script

```
#SBATCH --export=NONE
#SBATCH --partition=norm          Partition key says where to run
#SBATCH --job-name=AMBER16_MPI_T3
#SBATCH --output=res_amber16_t3.txt
#SBATCH --nodes=1
#SBATCH --ntasks=16
#SBATCH --time=18:10:00          Example: 18 hours and 10 minutes
#SBATCH --mem-per-cpu=10gb
```

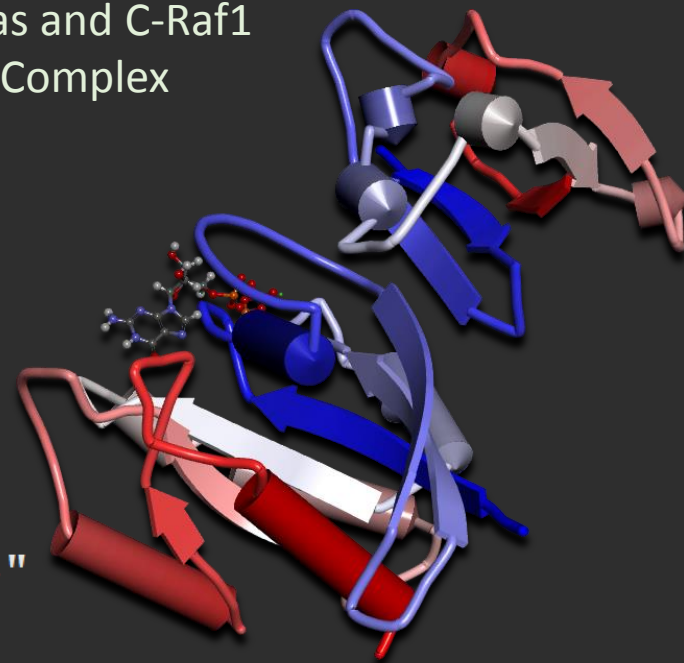
```
set NODE = "`hostname -s`"
module load amber
set AMBER_TUTORIAL_3_DATA = "$HOME/SLURM/Amber_SLURM/Tutorial_3"
# Please make sure the NCPUS match the "--ntasks" key value
set NCPUS = 16
```

```
setenv MYHOME /scratch/cluster_tmp/$USER/Amber16-Test
set P_SANDER = "$AMBERHOME/bin/sander.MPI"
cp $AMBER_TUTORIAL_3_DATA/* .
```

```
mpirun -n $NCPUS ${P_SANDER} -O -i min.in -o min.out -p ras-raf_solvated.prmtop \
      -c ras-raf_solvated.inpcrd -r min.rst -ref ras-raf_solvated.inpcrd
```

```
module unload amber
module unload intel_parallel
```

H-Ras and C-Raf1
Complex



Additional Information

- <https://slurm.schedmd.com/>
- HPC NIH
 - <https://hpc.nih.gov/docs/pbs2slurm.html>

Thanks

Ravi & Dennis