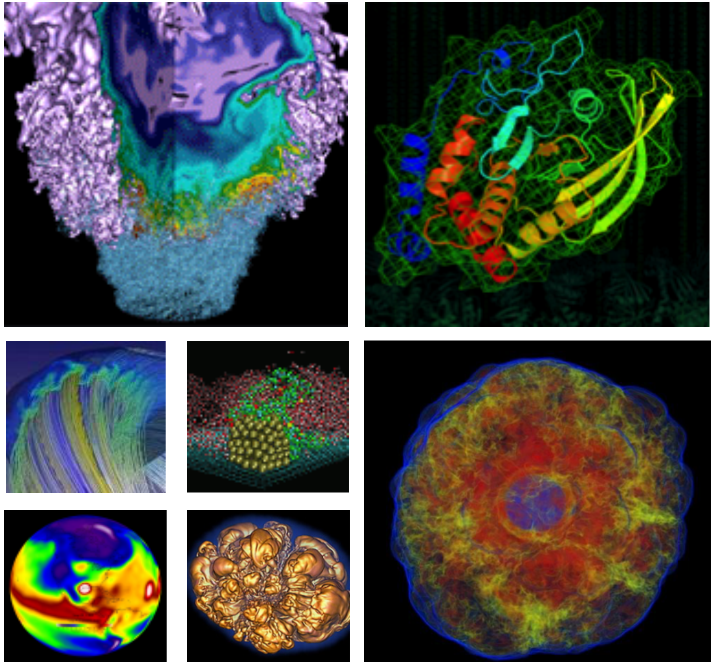
Slurm in Action: Batch Processing for the 21st Century



NERSC

Georg Rath

FOSDEM/04.02.2018

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

About NERSC

- Primary scientific computing facility of the office of science
- Two supercomputers, three clusters
  - Over 800.000 cores
  - Over 50 PB of storage in varying speeds
- Serving more than 6000 scientists
- Astrophysics, Climate & Earth Science, Chemistry, High Energy Physics, Genomics,…

What is Slurm?

- Formerly the **S**imple **L**inux **U**tility for **R**esource **M**anagement
- Highly scalable workload manager
- Runs >= 60% of TOP500 machines
- Development started 2001 at LLNL
- Commercial support by SchedMD
- Active community (over 150 contributors)

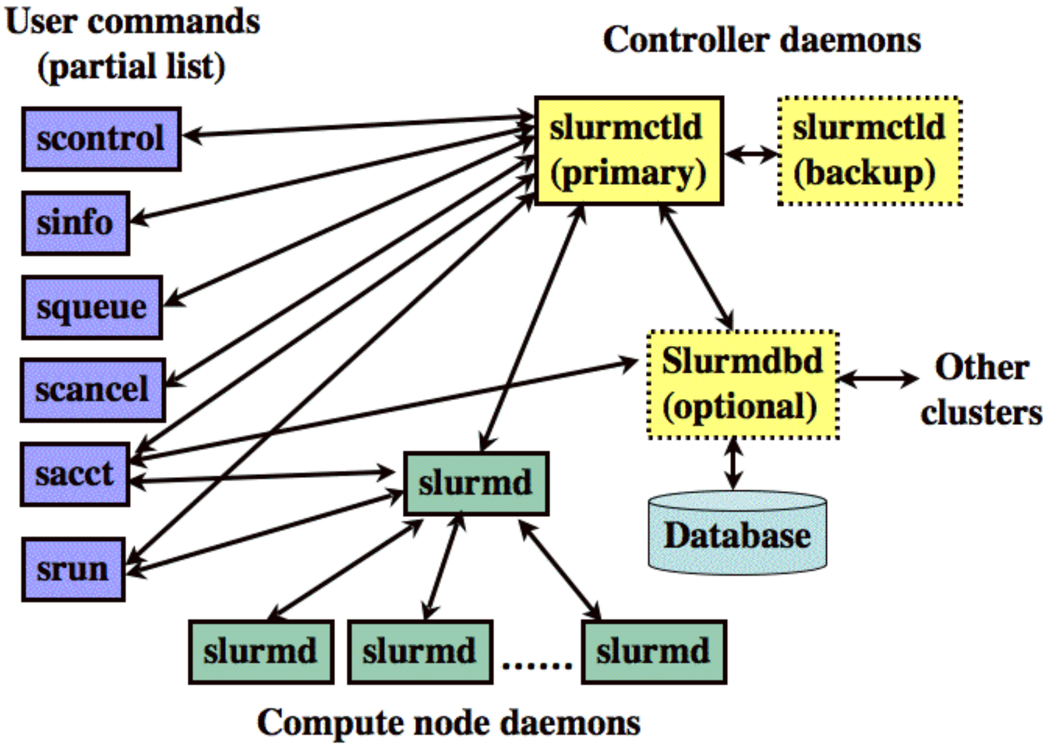
Basic Functionality of Slurm

- Write job script
- Submit to Slurm
- Job is queued and priority is applied
- Spawns the job on the requested resources when available
  - Enforces resource limits
  - Tracks resource usage
  - Takes care of cleanup when job finishes

Features

- Easy deployment
  - Single configuration file
  - Extensive documentation
- User friendly
- Highly scalable
- *Extremely* configurable

Architecture



Example Job Array

```
#!/bin/bash
#SBATCH --qos=high
#SBATCH --job-name=img_resize
#SBATCH --array=1-100
#SBATCH --cpus-per-task=2
#SBATCH --mem=10G

docker run image-preprocess --env INPUT=$SLURM_ARRAY_TASK_ID
```

Example Job Script with Dependecy

```
#!/bin/bash
#SBATCH --qos=high
#SBATCH --job-name=AI_job
#SBATCH --nodes=5
#SBATCH --dependency=afterok:procjobid

./ai-pipeline
```

Advanced Features

- Burst Buffers
- Container Integration
- Lua Job Submit Plugin
- Federation
- Plethora of Plugins

Burst Buffers

- Support for storage provisioning
- Initially developed for Cray DataWarp
  - But has a generic plugin that uses shell scripts
- Can automatically stage data

Example Burst Buffer

```
#!/bin/bash
#BB jobdw type=mybb capacity=1GB
#BB stage_in type=file source=s3:in destination=$JOB_BB/data
#DW stage_out type=file destination=s3:out source=$JOB_BB/data
docker run image-preprocess \
         --env INPUT=/home/g/data.in,OUTPUT=/home/g/data.out
```

Container Integration

- Realized via plugin
- Shifter, of course!
  - there also is an equivalent Singularity plugin
- Example:

```
salloc
    --image=custom:the_whole_stack:v3
```

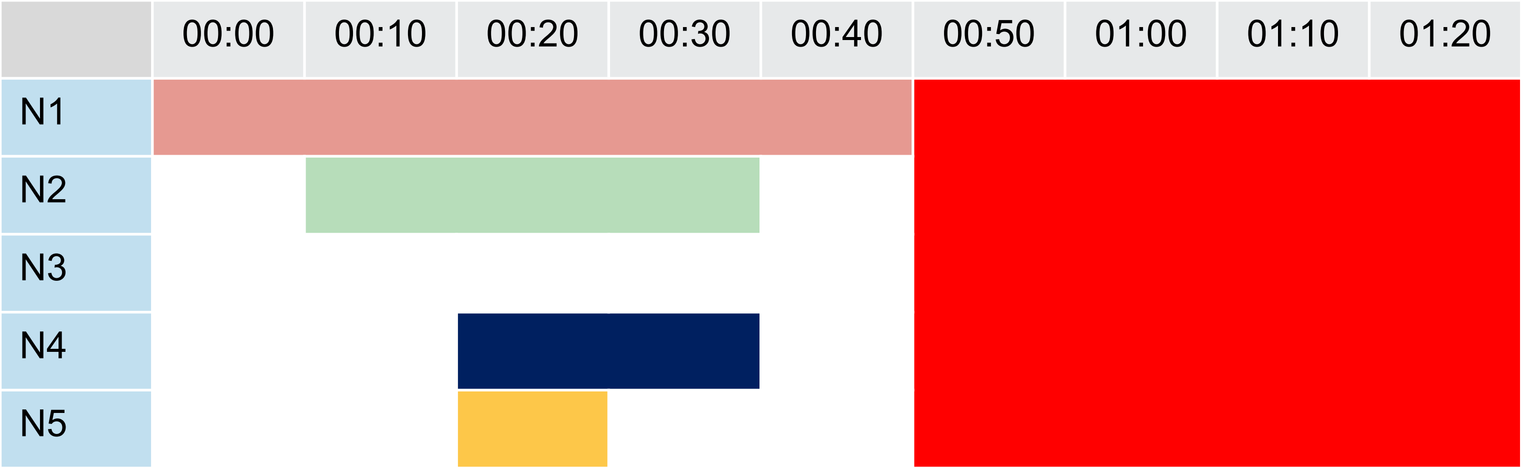
Lua Job Submit Plugin

- Execute a Lua file on job submit
- Can modify the job request
- *Very* powerful
  - can access Slurm internals

Plugins

- Like everything in Slurm: C or Lua (via LLNL wrapper)
- Lots of plugins floating around
  - X11
  - OOM notify
  - nersc-perf*

Queue with Backfill



Questions?

Thank You