Slurm

Recent Releases and Roadmap

Jacob Jenson
SchedMD

CUG 2018

Version 17.11

- Released November 2017
- Federated Clusters
- Heterogeneous Jobs
- Billing TRES

Version 17.11

- Federation
  - Scale out by scheduling multiple clusters as one
  - Submit and schedule jobs on multiple clusters
  - Unified views and jobid’s
  - Established through a central slurmdbd
  - Managed with sacctmgr command

Federation Capabilities

- **Job Distribution**
  - Jobs distributed across federation
  - Unique job IDs
- **Unified Views**
  - Appear as one cluster
- **Easy Administration**
  - Add/remove clusters to/from the federation with database commands

Unified Views

- Unified views provided with --federation command line option
  - Made default with FederationParameters=fed_display slurm.conf option
  - squeue, sinfo, sacct, sreport, sview etc.
  - --local, --clusters/-M options override federated view

```
$ export
SQUEUE_FORMAT2=jobarrayid:8,cluster:.8,statecompact:.4,origin:.8,siblingsviable:.16,siblingsactive:.16,timeu
sed:.8,numnodes:.6,nodelist:.12,reason:.15
$ squeue
JOBID    CLUSTER  ST  ORIGIN VIABLE_SIBLINGS ACTIVE_SIBLINGS    TIME NODES NODELIST      REASON
20132665    fed1  PD    fed3       fed1,fed3       fed1,fed3    0:00     5               Priority
67109269    fed1  PD    fed1  fed1,fed2,fed3  fed1,fed2,fed3    0:00     5               Resources
13421784    fed1  PD    fed2  fed1,fed2,fed3  fed1,fed2,fed3    0:00     5               Priority
20132665    fed3   R    fed3       fed1,fed3            fed3    2:44     5 fed3_[6-10]   None
13421784    fed3   R    fed2  fed1,fed2,fed3            fed3    2:47     5 fed3_[1-5]    None
20132665    fed2   R    fed3            fed2            fed2    2:50     5 fed2_[1-5]    None
67109268    fed2   R    fed1  fed1,fed2,fed3            fed2    2:50     5 fed2_[6-10]   None
13421783    fed1   R    fed2            fed1            fed1    2:54     5 fed1_[6-10]   None
67109267    fed1   R    fed1  fed1,fed2,fed3            fed1    2:57     5 fed1_[1-5]    None
```

Design Goals

- **Performance**
  - Little to no reduction in throughput of each cluster, performance scales with cluster count
- **Scalability**
  - No reduction in scalability of individual clusters
- **Ease of use**
  - Unified enterprise-wide view, minimize change in user interface
- **Stability**
  - No change in behavior for clusters not explicitly placed into a federation

Configuration

- A cluster can only be part of one federation at a time
- Jobs can’t span clusters

Persistent Connections

- Clusters talk to each other over persistent connections
  - Reduces communication overhead -- only authenticate once
  - Broken connections detected immediately and established when needed
  - Controller and SlurmDBD use the same code

Job Submission

- sbatch, salloc, srun supported
- Jobs submitted to local cluster
- Sibling jobs submitted to all “viable” clusters
  - viable == all clusters ||
  - --clusters=<clusters> & --cluster_constraint=<features>
- Job stays on the local cluster -- even if not viable -- to coordinate and route requests to/from sibling clusters
  - Job starts, updates, cancellations

Scheduling

- Federated jobs contain the locations of all “sibling” jobs
- Each cluster independently schedules each sibling job
- Coordinates with “origin” cluster to start job
  - The origin cluster is determined from the job id
  - Prevents multiple jobs from being started at the same time
  - Policies in place to handle if origin cluster fails
- Once sibling job is started, origin cluster revokes remaining siblings jobs
- Batch jobs can be requeued to federation

Heterogenous Jobs

- Join resource allocation requests into a single job.
- As an example, this makes it easy to allocate a job with 10 Haswell nodes and 1000 KNL nodes.
    - Currently, this is difficult to accomplish, and requires careful manipulation of --constraint and CPU count calculation.

Submitting Hetereogenous Jobs

- Multiple independent job specifications identified in command line using “:” separator
- The job specifications are sent to slurmctld daemon as a list in a single RPC
- The entire request is validated and accepted or rejected
- Response is also a list of data (e.g. job IDs)

```
$ salloc -n1 -C haswell : -n256 -C knl bash
```

Heterogeneous Batch Jobs

- Job components specified using “:” command line separator OR
- Use “#SBATCH” options in script separating components using “#SBATCH packjob”
- Script runs on first component specified

```
$ echo my.bash
#!/bin/bash
#SBATCH -n1 -C haswell
#SBATCH packjob
#SBATCH -n256 -C knl
…
$ sbatch my.bash
```

Billing TRES

- New “billing” TRES
  - On by default -- AccountingStorageTRES
  - Enforce limits on usage calculated from partition’s TRESBillingWeights
  - Use existing limits (GrpTRESMins, GrpTRESRunMins, GrpTRES, MaxTRESMins, MaxTRES, etc.)
  - Usage seen with scontrol show jobs, sacct, sreport.

Version 18.08

- Release scheduled for August 2018
- Google Cloud support (integration scripts provided)
- Support for MPI jobs that span heterogeneous job allocations
- Support for multiple backup slurmctlds
- Improvements to KNL scheduling and CPU binding
- Cray
    - Manage persistant DataWarp allocations without allocating compute nodes. (“--nodes=0”)
    - “scontrol show dwstat” - report output from ‘dwstat’ command

and Beyond!

- cons_tres
  - First step in replacing cons_res
  - Enable Generic Resources (GRES) to be scheduled backfilled just like CPUs
    - Focus for first release will be for improved GPU scheduling
  - Job commands will be updated with new options

Questions