Slurm 20.02 Roadmap and Beyond

Tim Wickberg
SchedMD

Slurm User Group Meeting 2019

Slurm Release Schedule

- Slurm major releases come out every nine months
- Major release numbers are the two digit year, period, two digit month
  - 19.05 ⇒ 2019, May
- Maintenance releases, such as 19.05.3, come out roughly monthly for the most recent major release
- Two most recent major releases are still supported
  - This is 19.05 and 18.08 currently

Upcoming Slurm Releases

- 20.02 - February 2020
- 20.11 - November 2020

REST API

- See separate presentation
- Initial version handles common slurmctld interactions

AuthAltTypes

- Allow slurmctld to talk different authentication protocols simultaneously
  - Added in 19.05, but the only other auth plugin is auth/none
    - Which has zero security, and should never run in production
- Add a new auth/token plugin
  - Tokens will be managed through slurmdbd / sacctmgr
  - Two types initially:
    - Global, permitted to submit and manage jobs as any user
    - User, permitted for only one user account

"Config-less" Slurm

- New way to setup the cluster
  - Meant to handle systems where pushing config file updates out is difficult
    - E.g., 20k nodes take 40+ minutes to sync the caching filesystem
  - Only need configuration file on slurmctld hosts
- Add DNS SRV records for the slurmctld hosts
- slurmd processes will look this up and fetch the config info from an RPC call to slurmctld if /etc/slurm.conf does not exist
  - Alternative to DNS: specify slurmctld by a new CLI option to slurmd

"Config-less" Slurm

- User commands need some work to avoid overloading slurmctld
- slurmd will expose a UNIX socket as /run/slurm.sock
- Command-line options:
  a. use /etc/slurm/slurm.conf if available
  b. try local unix socket at /run/slurm.conf to slurmd
  c. or fall back DNS SRV record otherwise for config
     - needed for login nodes

Retroactive WCKey Updates

- "sacctmgr update jobid=<foo> set newwckey=correctkey"
- Supports selection by user, current wckey, and can limit to a specific date range
- Rerolls usage so sreport data is updated as well

OverSubscribe=EXCLUSIVE

- Update OverSubscribe=EXCLUSIVE to always assign all TRES in the job to the job
  - OverSubscribe=EXCLUSIVE is used to always provide full-node allocations on a partition
  - Current (Slurm <= 19.05) behavior is to assign all CPUs and Memory, but not to assign any further GRES automatically

FastSchedule is gone

- Remove FastSchedule option
  - FastSchedule=0 does not work properly with cons_tres
  - Deprecated in 19.05.3+, you will see errors in slurmctld/slurmd log files warning about this
- New SlurmdParameters=config_overrides
  - Replaces FastSchedule=2 functionality
  - Used for test/development when you need to lie about the actual hardware
  - Still not recommended for production use

burst_buffer/datawarp additions

- Adding % replacement syntax for #DW / #BB directives
- Replace the symbol with the correct value for the job

#DW / #BB Symbol	Replacement
\\	Stop further symbol processing.
%%	A single % symbol.
%A	Job array's master job allocation number.
%a	Job array ID (index) number.
%d	WorkDir.
%j	Job ID.
%u	User name.
%x	Job name.

Prolog/Epilog Refactoring

- Move Prolog/Epilog/PrologSlurmctld/EpilogSlurmctld behind a new plugin interface
  - Current script functionality moves into the "script" plugin type
  - Allows easier access to the underlying job launch data
  - If you have a good name for this plugin type, I haven't found a good name - "ProEpiLogInterfacePlugin" is a bit unwieldy
  - Make the environment variables more consistent across each interface

Adjustments to PMI

- Change how libpmi.so (PMI1) links to avoid direct dependency on libslurm.so.<VERSION>
- Workaround for OpenMPI statically linking to our libpmi.so, and thus inheriting a dependency on libslurm.so.<VERSION>
  - Which then breaks your OpenMPI installs for each Slurm upgrade

Packaging

- slurm.spec-legacy has been removed
  - We've warned about this since before 17.11

Slurm 20.11 Roadmap

REST API

- Extend to cover common slurmdbd interactions

Heterogeneous Job Steps

- Similar to HetJobs, but extended to step launch within an existing "normal" job

Reservation Affinity

- Add "Promiscuous" option to Reservations
- Jobs with matching account/qos settings will be eligible to run in these reservations even if they have not specified --reservation on the submission
  - They will still be considered for execution outside of the reservation

TRES

- Add new --ntasks-per-gpu option
  - Does what it says on the tin

Expose Additional Scheduling Details

- Mark nodes blocked from running jobs by a future larger job as something other than IDLE
  - Exact display name still TBD (e.g. PreAllocated)
  - Accounting will still reflect these nodes as IDLE, but at least sinfo will separate them and keep your users from complaining that their job won't launch while nodes are IDLE
- Expose a timestamp of the last backfill cycle to consider the job for execution
  - Useful for backfill tuning

Slurm 20.02 Anti-Roadmap

Pending code removals

- gres/mic plugin for Intel KNC coprocessor cards
  - And associated mpirun-mic script

Pending code removals

- smap command
  - After removing all the BlueGene and Cray/ALPS code, this is really just an odd version of squeue

Pending code removals

- "Layouts / Entities"
  - Is anyone using this in production?
  - I've seen some sites with it half-configured… not any working examples

Questions?