Slurm 20.11 and Beyond
Open Q+A

Tim Wickberg
SchedMD

Slurm User Group Meeting 2020

Agenda

All times are US Mountain Daylight (UTC-6)

Time	Speaker	Title
9:00 - 9:50	Jason Booth	Field Notes 4: From The Frontlines of Slurm Support
10:00 - 10:25	Brian Christiansen	Cloud and Stuff
10:30 - 10:50	Nate Rini	REST API
11:00 - 11:50	Tim Wickberg	Slurm 20.11 and Beyond, Open Q+A

Welcome

- Four separate presentations, four separate streams
- Presentations will remain available for one week after SLUG'20 concludes
- Presentations are available through the SchedMD Slurm YouTube channel
  - https://youtube.com/c/schedmdslurm
- Or through direct links from the agenda
  - https://slurm.schedmd.com/slurm_ug_agenda.html

Asking questions

- Feel free to ask questions throughout through YouTube's chat
- Chat is moderated by SchedMD
- Questions will be relayed to the presenter by the moderators
  - Some may be deferred to the end if they cannot be relayed in a timely fashion
- Note there is a ~5 second broadcast delay
  - By the time you've asked your question, the presenter may have moved ahead to a different topic, and may defer the question until the end

Slurm 20.11 and Beyond

Tim Wickberg
SchedMD

Slurm Releases

- 20.02 - February 2020
- 20.11 - November 2020
- 21.08 - August 2021

Slurm Release Schedule

- Slurm major releases come out every nine months
- Major release numbers are the two digit year, period, two digit month
  - 20.02 ⇒ 2020, February
- Maintenance releases, such as 20.02.5, come out roughly monthly for the most recent major release
- Two most recent major releases are still supported
  - This is 20.02 and 19.05 currently

Slurm 20.02 Release

Tim Wickberg
SchedMD

REST API

- See separate presentation
- Initial version handles common slurmctld interactions

AuthAltTypes

- Allow slurmctld to talk different authentication protocols simultaneously
- Add a new auth/jwt plugin
  - Users can request tokens through `scontrol token`

Configless Slurm

- New way to setup the cluster
- See examples from "Field Notes 4" earlier today

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 - "PrEpPlugins"
  - Current script functionality moves into the "script" plugin type
  - Allows easier access to the underlying job launch data

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

NodeSet syntax for slurm.conf

- New "NodeSet" syntax for slurm.conf
- Define a NodeSet as a set of Nodes
  - Or, select all nodes with a given Feature defined
- And then use the node sets interchangeably with the node names as part of your Partition definitions
  - Rather than error-prone copy+pasting long lists

NodeSet syntax for slurm.conf

```
NodeName=node[0001-0005] Features=e2620v4,xeon,qdr,v100
NodeName=node[0006-0010] Features=e2620v5,xeon,qdr
NodeName=node[0011-0015] Features=e2620v6,xeon,ndr

NodeSet=v100nodes Feature=v100
NodeSet=random Nodes=node[0001,0003,0008-0013]
NodeSet=imaginary Feature=e2620v6

PartitionName=v100 Nodes=v100nodes
# -> node[0001-0005]
# these two sets overlap in part, but the slurmctld would de-duplicate for us:
PartitionName=somestuff Nodes=random,imaginary
# -> node[0001,0003,0008-0015]
```

"Magnetic" Reservations

- Add "Magnetic" 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

Slurm 20.11 Roadmap

REST API

- Extend to cover common slurmdbd interactions

IPv6

- IPv6 support throughout Slurm
- For 20.11, Slurm will require a CommunicationParameters option to enable dual-stack support
- For 21.08 (tenatively), we will enable dual-stack by default, with an option to force IPv4-only

MariaDB SSL Connection Support

- See StorageParameters in slurmdbd.conf(5) for setup

Heterogeneous Job Steps

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

Heterogeneous Job Steps

```
tim@blackhole:~$ salloc -N 2 --exclusive
salloc: Granted job allocation 24217
tim@blackhole:~$ srun -N 1 echo a : -N 1 echo b
a
b
tim@blackhole:~/slurm$ sacct -j 24217
       JobID    JobName  Partition    Account  AllocCPUS      State ExitCode
------------ ---------- ---------- ---------- ---------- ---------- --------
24217              bash    general       root          8    RUNNING      0:0
24217.extern     extern                  root          8    RUNNING      0:0
24217.0+0          echo                  root          1  COMPLETED      0:0
24217.0+1          echo                  root          1  COMPLETED      0:0
```

--threads-per-core

- Previously only affected allocation
- Now influences placement of tasks
- Implies `--cpu-bind=threads`
- Like `--hint=nomultithread`, but more control for threads>2

--threads-per-core

- Max number of threads per core
  - `srun -n1 -c2 --threads-per-core=2 prog`
    - Places two cpus on one 2 threaded core
  - `srun -n1 -c2 --threads-per-core=1 prog`
    - Places one cpu per core

"Interactive" Job Step

- Easier way to force a user's terminal to the compute nodes when using salloc
- Replace complicated SallocDefaultCommand settings with a new "Interactive" step
- Tracked in accounting appropriately, and will not cause confusing step-launch issues for GRES or GPUs

TRES

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

Mail Type

- New "Invalid Dependency" mail type
  - Message sent when job is removed due to invalid dependencies
    - Due to DependencyParameters=kill_invalid_depend

Reservations

- Allow users to delete reservations
  - Enable with new SlurmctldParameters=user_resv_delete option
  - Only permitted if they would have been allowed to run within the reservation
- Allow multi-reservation job submission:
  - `sbatch --reservation=foo,bar,baz myjob.sh`

Reservations

- New AllowGroups access control on a reservation
  - Permit access by UNIX group
- Skip the next occurance of a repeating reservation:
  - `scontrol update reservation=weekly_resv skip`

scrontab

- Permit users to submit recurring jobs, with a crontab compatible syntax for recurrence
- And add a new "scrontab" user command to manage them

Slurm 20.11 Anti-Roadmap

Code removals

- "Layouts / Entities"
  - Finally removed
- Message Aggregation
  - Removed in favor of new method for RPC queuing in slurmctld

… and Beyond

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

HPE Cray Shasta Support

- In progress

Upcoming Events

SC20 Conference

- There will be a Slurm Community BoF at the virtual SC20
  - Details to be announced, will be posted to slurm-announce and slurm-users when available
  - Similar-but-even-more-condensed format to SLUG'20

SLUG'21

- While we'd usually announce our next year's location at the end of SLUG, unfortunately we cannot commit at the moment
- Expect further details (through slurm-announce and slurm-users mailing lists) in the spring

SchedMD is Hiring

- SchedMD is always looking to hire experienced systems programmers and support staff
- https://schedmd.com/careers.php
- jobs@schedmd.com

Open Q+A

End Of Stream

- Thanks for watching!