Slurm State of the Union
SC16

Tim Wickberg - SchedMD

Code statistics

- Over 128 contributors
- 570 kloc on master branch
  - slurm/src directory only, excludes contribs
  - Additional 120kloc in testsuite/expect for regression tests
- In the past year
  - 1685 files changed, 99483 insertions(+), 174799 deletions(-)
  - 2782 commits
  - 63 contributors

Code contribution

Please submit patches through https://bugs.schedmd.com

See new CONTRIBUTING.md file for code style advice and code submission workflow.

16.05 - Releases

- 16.05.0 Released in May 2016
- Current point release: 16.05.6
- Next major release: 17.02
  - Still on a nine-month release schedule

16.05 - Partition Options

- Partition option “Priority” split to two new settings:
  - "PriorityTier" - only affects preemption.
  - "PriorityJobFactor" - only affects job priority.
    - PriorityWeightPartition now useful.

Partition: A	
Priority:	2
Jobs	Priority
123	200
124	150

Partition: B	
Priority:	1
Jobs	Priority
125	400
126	350

16.05 - Partition Options

- Partition option "Shared" renamed to "OverSubscribe"
  - "--shared" option to salloc/sbatch/srun changed to "--oversubscribe".

16.05 - Cgroups

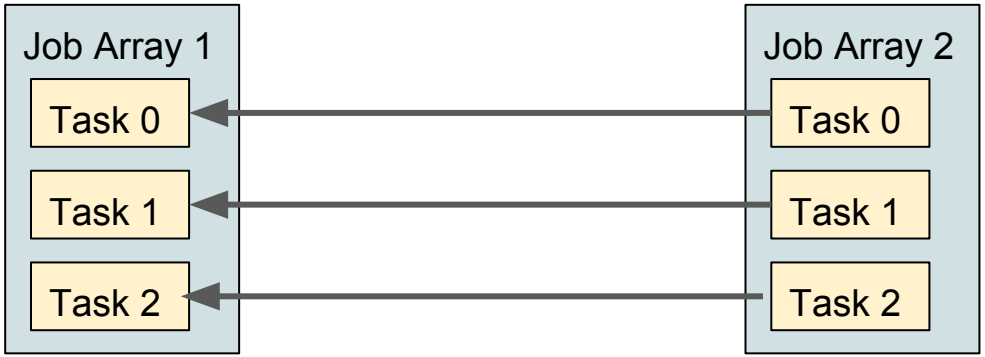
- Change default CgroupMountpoint (in cgroup.conf) from "/cgroup" to "/sys/fs/cgroup" to match current standard.
- 16.05.5+ does not require the ReleaseAgent setting.
  - All cgroup hierarchies should be cleaned up on job completion by slurmstepd.
  - Avoids mount option conflict with systemd.

16.05 - Highlights

- PMIx protocol support added.
- Add Multi-Category Security (MCS) infrastructure to permit nodes to be bound to specific users or groups.
- Added --deadline option to salloc, sbatch and srun. Jobs which can not be completed by the user specified deadline will be canceled with a state of "Deadline" or "DL".

16.05 - Job Array Dependencies

- Added new job dependency type of "aftercorr" which will start a task of a job array after the corresponding task of another job array completes.

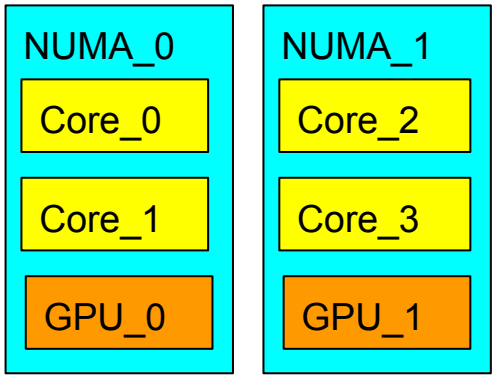


16.05 - Wrappers for other WLMs

- Added wrappers for LSF/OpenLava commands.
- Added Grid Engine options to qsub command wrapper.

16.05 - GRES Binding

- Add --gres-flags=enforce-binding option to salloc, sbatch and srun commands
  - Matching socket/NUMA required for CPU and GRES allocations



GPU_0 can only be allocated with Core_0 and/or Core_1

GPU_1 can only be allocated with Core_2 and/or Core_3

16.05 - KNL

- Added node_features plugin infrastructure
- Split node Features field into two parts
  - Active features
  - Available features (can become active after reboot)
- Added support for KNL on Cray systems with 16.05.0
- Added node_features/knl_generic plugin in 16.05.6
  - Support for KNL mode setting on generic Linux systems

17.02

- February 2017
- Federation
  - Scheduled to be fully complete by May 2017 (Slurm 17.11).
  - Separate presentation Wednesday at 1.
- Support for >= 2TB memory per node.
  - If you have custom plugins, make sure to convert all memory variables to uint64_t from uint32_t.

17.02

- New sacctmgr commands:
  - "shutdown": shutdown slurmdbd
  - "list stats": get slurmdbd statistics - “sdiag for slurmdbd”
  - "clear stats": clear slurmdbd statistics.
- New MailDomain setting to qualify usernames.
  - Avoid MTA configuration issues on controller.

17.02

- New salloc/sbatch/srun “--spread-job” option to distribute tasks over as many nodes as possible.
  - Treats the --ntasks-per-node option as a maximum value.
- New “AdminComment” field.
  - Like “Comment”, but only modifiable by admins.

17.02 Code Cleanup

- Remove AIX support.
- BlueGene/L and /P support removed
- Lots of additional code cleanup.
- Code base moving to full C99 and POSIX1.2008 compliance.

17.02

- The database index for jobs is now 64-bits
  - Was 32-bits before
    - Limit of 4-billion jobs stored in the database, now 18-quintillion.
- If you happen to be close to 4 billion jobs in your database, you will want to update your slurmctld at the same time as your slurmdbd to prevent roll over of this variable.

17.02 - Warnings

- **17.02 is in development. Don’t use for production.**
- **The database schema has changed. Updating slurmdbd will take time. No records will be lost while upgrading, but the slurmdbd may not be responsive. It will not be possible to automatically revert the database to an earlier version of Slurm**
- The max MaxJobID is now 67,108,863. Any pre-existing jobs will continue to run but new job ids will be within the new MaxJobID range. Adjust your configured MaxJobID value as needed to eliminate any confusion.
- Every plugin except SPANK must be built against the same version of Slurm (major and minor version number) to be loaded

Infinity and Beyond

- Retirement of Cray ALPS support.
  - Native Cray offers significant advantages and access to new functionality.
- Retirement of sched/wiki and sched/wiki2 interfaces.
- BlueGene/Q retirement.
  - Last supported release likely ~ 17.11 or 18.08.

SchedMD

- SchedMD main office moved to Lehi, UT, over the summer.
- We’re hiring!
  - https://www.schedmd.com/careers.php