



# Slurm Birds of a Feather


Tim Wickberg  
SchedMD

SC17

# Outline



- Welcome
- Roadmap
  - Review of 17.02 release (February 2017)
  - Overview of upcoming 17.11 (November 2017) release
  - Roadmap for 18.08 and beyond
- Time Remaining - Open Q+A



# Slurm 17.02

Copyright 2017 SchedMD LLC  
<https://www.schedmd.com>

# Version 17.02

- Released February 2017
- Some background restructuring for Federation
- Relatively few changes visible to users or admins
- 1,913 commits ahead of 16.05 release.
  - 1948 files changed, 48407 insertions(+), 62333 deletions(-)

# sbcast improvements



- sbcast - introduced in 2006, uses the hierarchical communication to propagate files to compute nodes.
- Added `srun --bcast` to fan out command binary as part of launch.
- Added lz4 and gzip compression options.
  - Set through `--compress` option, or `SBCAST_COMPRESS`, and/or *SbcastParameters* option in `slurm.conf`.
  - lz4 highly recommended. gzip not recommended in most environments.

# slurmdbd daemon statistics

- sdiag, but for slurmdbd
  - `sacctmgr show stats` - Reports current daemon statistics
  - `sacctmgr clear stats` - Clear daemon statistics
  - `sacctmgr shutdown` - Shutdown the daemon

# slurmdbd daemon statistics

```
$ sacctmgr show stats
```

```
Rollup statistics
```

Hour	count:8	ave_time:150348	max_time:342905	total_time:1202785
Day	count:1	ave_time:285012	max_time:285012	total_time:285012
Month	count:0	ave_time:0	max_time:0	total_time:0

```
Remote Procedure Call statistics by message type
```

DBD_NODE_STATE	( 1432)	count:40	ave_time:979	total_time:39162
DBD_GET_QOS	( 1448)	count:12	ave_time:949	total_time:11389
...				

```
Remote Procedure Call statistics by user
```

alex	( 1001)	count:18	ave_time:1342	total_time:24156
...				

# Other 17.02 Changes



- Cgroup containers automatically cleaned up after steps complete
- Added *MailDomain* configuration parameter to qualify email addresses
- Added *PrologFlags=Serial* configuration parameter to prevent Epilog from starting before Prolog completes (even if job cancelled while Prolog is active)



# Other 17.02 Changes



- Added burst buffer support for job arrays
- Memory values changed from 32-bit to 64-bit, increasing maximum supported limit enforcement and schedule for nodes above 2TB
- Removed AIX, BlueGene/L, BlueGene/P, and Cygwin support
- Removed sched/wiki and sched/wiki2 plugins



# Slurm 17.11

Copyright 2017 SchedMD LLC  
<https://www.schedmd.com>

# Version 17.11



- To be released in November 2017
- Two big ticket items:
  - Federated Clusters support
  - Heterogeneous Job support
- 2,359+ commits ahead of 17.02
  - 1227 files changed, 82452 insertions(+), 53719 deletions(-)
  - Release candidates out now, 17.11.0 release at the end of November.

# Version 17.11



- Federation
  - Scale out by scheduling multiple clusters as one
  - Submit and schedule jobs on multiple clusters
  - Unified jobid's
  - Unified views
  - Established through a central slurmdbd, managed with sacctmgr command.
- For more details please see the Federation presentation.

# 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.

# Heterogenous 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
```

# Submitting Heterogeneous 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
```



# Heterogeneous Job Data Structure



- Each component of a heterogeneous job has its own job structure entry
- “Head” job has pointers to all components (like job arrays)
- New fields
  - JobID - Unique for each component of the heterogeneous job
  - PackJobID - Common value for all components
  - PackJobOffset - Unique for each component, zero origin
  - PackJobIdSet - List of all job IDs in the heterogeneous job

# Heterogenous Job Management

- Standard format ID for managing heterogeneous jobs is “<PackJobID>+<PackJobOffset>”

```
$ squeue --job=93
```

JOBID	PARTITION	NAME	USER	ST	TIME	NODES	NODELIST
93+0	debug	test	adam	R	4:56	1	nid00001
93+1	debug	test	adam	R	4:56	2	nid000[10-11]
93+2	debug	test	adam	R	4:56	4	nid000[20-23]

# Version 17.11

- Changed to dynamic linking by default
  - Single libslurmfull.so used by all binaries.
  - Install footprint drops from 180MB (17.02) for {bin, lib, sbin}, to 83MB.
  - Use `--without-shared-libslurm` configure option to revert to old behavior.
- Native Cray is now the default for `--with-cray`

# Version 17.11

- Overhaul of slurm.spec
  - Rearrangement components into obvious packages
    - slurm - libraries and all commands
    - slurmd - binary + service file for compute node
    - slurmctld - binary + service file for controller
    - slurmdbd - binary, mysql plugin, + service file for database
  - “Correct” support for systemd service file installation.
    - No SystemV init support, assumes RHEL7+ / SuSE12+ environment.
  - Older version preserved as contribs/slurm.spec-legacy
    - Deprecated, and will receive minimal maintenance

# Version 17.11

- Removal of Solaris support
- Removal of obsolete MPI plugins
  - Only PMI2, PMIx, and OpenMPI/None remain.
  - The PMI2 and PMIx APIs are supported by all modern MPI stacks.
  - And almost all can launch without those APIs by interpreting SLURM\_\* environment variables directly.
  - Note that the OpenMPI plugin is identical to the None plugin.
    - MpiDefault=openmpi doesn't do anything.
    - Please update your configs, it will be removed in a future release.

# Version 17.11

- AccountingStorage=MySQL no longer supported in slurm.conf
  - Mode that used to allow you to get minimal accounting support without slurmdbd.
  - Has not worked correctly for some time.
  - Use AccountingStorage=SlurmDBD in slurm.conf.
    - Along with AccountingStorage=MySQL in slurmdbd.conf

# Version 17.11

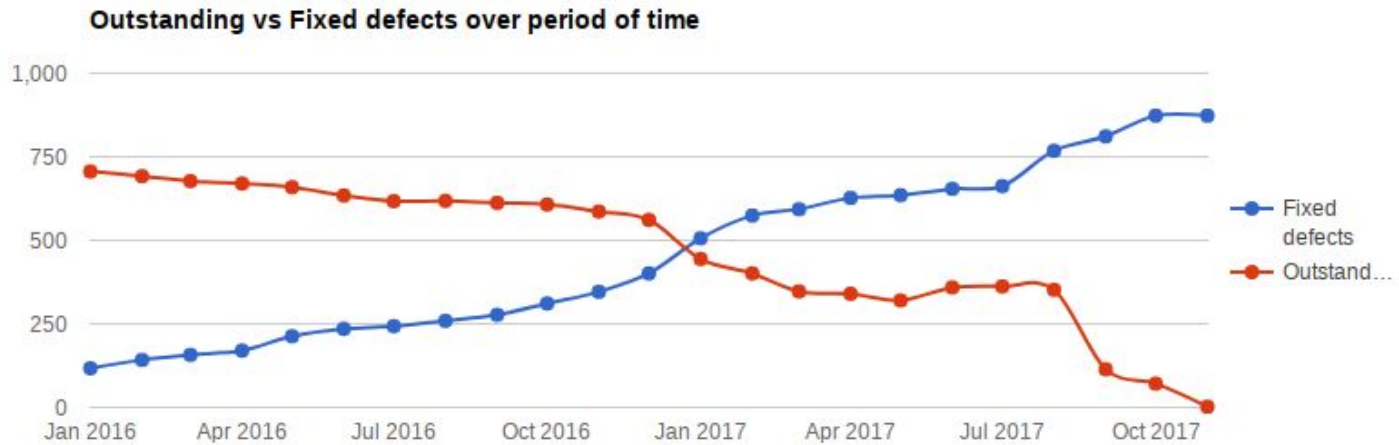
- Log rotation: SIGUSR2 to {slurmctld, slurmdbd, slurmd} will cause it to drop and reopen the log files.
  - Use instead of SIGHUP, which accomplishes this by reconfiguring the daemon. (Which will cause re-registration, and can cause performance issues in slurmctld.)
- MaxQueryTimeRange in slurmdbd.conf.
  - Limit range in a single sacct query, helps prevent slurmdbd hanging/crashing by trying to return too large of a data set.

# Version 17.11

- Additional resiliency in slurmd high-availability code paths.
  - Avoids split-brain issues if the path to StateSaveLocation is not the same as the network path to the nodes.
- Code hardening using Coverity
  - Static analysis tool, free for open-source projects
  - Outstanding issue count reduced from 725 to 0
  - This compliments the long-standing use of `-Wall` on all development builds, and testing using several compilers and platforms.



# Coverity



# Built-in X11 Forwarding

- Similar functionality to CEA's SPANK X11 plugin.
  - Use configure option of `--without-x11` to continue using SPANK plugin instead.
- Implementation uses libssh2 to setup and coordinate tunnels directly.
- Adds `--x11` option to `salloc/sbatch/srun`.
  - Optional arguments control which nodes will establish tunnels:  
`--x11={all, batch, first, last}`

# Built-in X11 Forwarding



- Enable with *PrologFlags=X11*
  - *PrologFlags=Contain* is implied.
  - Uses the “extern” step on the allocated compute node(s) to launch one tunnel per node, regardless of how many steps are running simultaneously.
- Users must have either SSH hostkey authentication or password-less SSH keys installed inside the cluster.
- Can work alongside `pam_slurm_adopt` to set the correct `DISPLAY` on SSH processes when forwarding is in place.

# Centralized Extended GID lookup



- Lookup the extended GIDs for the user before launching the job, and send as part of the job credential to all allocated slurmd's.
- Enable with *LauchParameters=send\_gids*

# Centralized Extended GID lookup



- Avoids compute nodes all making simultaneous calls to `getgrouplist()`, which has been a scalability issue for  $>O(1000)$  nodes.
  - If `sssd/ldap` fails to respond promptly, `getgrouplist()` may return with no extended groups.
  - Leaving a job with a mix of nodes with and without the correct gids.


# 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.

# Other 17.11 Changes



- More flexible advanced reservations (FLEX flag on reservations)
  - Jobs able to use resources inside and outside of the reservation
  - Jobs able to start before and end after the reservation
- `sprio` command reports information for every partition associated with a job rather than just one partition
- Support for stacking different interconnect plugins (JobAcctGather)
- Add `scancel --hurry` option
  - Cancel job without staging-out burst buffer files
- `sdiag` reports DBD agent queue size



# Slurm 18.08

Copyright 2017 SchedMD LLC  
<https://www.schedmd.com>



# Version 18.08



- Release 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 DataWarp allocations without allocating compute nodes. (“--nodes=0”)
  - “scontrol show dwstat” - report output from ‘dwstat’ command

# Version 18.08



- A few planned anti-features:
  - Remove support for Cray/ALPS mode
    - Must use native Slurm mode (recommended for some time)
  - Remove support for BlueGene/Q
  - Remove or repair support for macOS
    - Has been broken for years due to linking issues
    - Patch submissions welcome

# and Beyond!

