Slurm Version 18.08 Overview

Brian Christiansen
SchedMD

Slurm User Group Meeting 2018

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Schedule

- Previous major release was 17.11 (November 2017)
- Latest major release 18.08 (August 2018)
- Next major release planned 19.05 (May 2019)

Copyright 2018 SchedMD LLC
http://www.schedmd.com

18.08 Highlights

- Heterogeneous environments
- Burst buffer enhancements
- Fault tolerance
- Cloud computing
- Queue stuffing
- New TRes
- New TRes reporting options
- And more...

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Heterogeneous Job Steps (MPI)

- A single MPI_COMM_WORLD can span multiple heterogeneous job components
  - Can vary CPUs, memory, GPUs, etc. for job component
- Each component of the job will internally have a separate step record
- MPI will be presented with a view of one single job, spanning all Slurm job components, using environment variables
- Not compatible with Cray MPI or ALPS libraries
- sacct can report all components with “--whole-hetjob=yes” option

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Heterogeneous Features Specifications

- Applies to both advanced reservations and job requests
- Adds support for parenthesis and counts
- Favors use of nodes with features currently available
  - Avoid KNL reboot for requested MCDRAM or NUMA mode

```
$ sbatch --constraint=[(knl&snc4&flat)*4&haswell*1]    ....
```

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Heterogeneous Features (continued)

- Job can specify required features of batch host

```
$ sbatch --constraint=[(knl&snc4&flat)*4&haswell*1]  --batch=haswell  ....
```

Copyright 2018 SchedMD LLC
http://www.schedmd.com

KNL Enhancements

- Added node and partition “CpuBind” configuration parameter to control default task binding (in slurm.conf)
- Added “NumaCpuBind” configuration parameter to knl.conf
  - Changes node’s default CPU bind mode based upon KNL NUMA mode
- Added “ValidateMode” to knl.conf for static KNL configurations
- Added “NodeRebootWeight” configuration parameter to knl.conf
- Report node feature plugin configuration with “scontrol show config”

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Burst Buffers

- Added “scontrol show dwstat” command to report burst buffer status
    - Added “GetSysStatus” field to burst_buffer.conf file
- Added new job state of SO/STAGE_OUT while stage-out in progress
- Added new burst buffer state of “teardown-fail”
- Burst buffer errors logged to new job “SystemComment” field
- Enable jobs with zero node count for creation or deletion of persistent burst buffers
- Add “SetExecHost” flag to burst_buffer.conf to enable access from login node for interactive jobs

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Arbitrary Count of slurmctld backups

- Configuration parameters ControlMachine, BackupController, etc replaced with ordered list of hostnames and IP addresses
  - Old options are still supported
- Added SlurmctldPrimaryOnProg and SlurmctldPrimaryOffProg configuration options to run scripts when primary/backup changes
  - Scripts can change IP routing if desired

```
# Excerpt of slurm.conf file
SlurmctldHost=head1(12.34.56.78)
SlurmctldHost=head2(12.34.56.79)
SlurmctldHost=head3(12.34.56.80)
SlurmctldAddr=10.120.10.1
SlurmctldPrimaryOnProg=/opt/slurm/default/etc/primary_on
SlurmctldPrimaryOffProg=/opt/slurm/default/etc/primary_off
```

Cloud Enhancements

- ResumeFailProgram
  - The program that will be executed when nodes fail to resume by ResumeTimeout. The argument to the program will be the names of the failed nodes (using Slurm's hostlist expression format).

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Queue Stuffing

- New association/qos options
- GrpJobsAccrue=<max jobs>
  - Maximum number of pending jobs in aggregate able to accrue age priority for this association and all associations which are children of this association.
- MaxJobsAccrue=<max jobs>
  - Maximum number of pending jobs able to accrue age priority at any given time for the given association. This is overridden if set directly on a user. Default is the cluster's limit.

Copyright 2018 SchedMD LLC
http://www.schedmd.com

New TRes

- New TRES
  - fs/disk, fs/lustre, ic/ofed, vmem, pages
- New default TRES
  - cpu, mem, energy, node, billing, **fs/disk, vmem, pages**
- AcctGather{FileSystem|Infiniband}Type
  - Not just for profiling anymore.
  - Must define fs/lustre and/or ic/ofed in AccountingStorageTRES

Copyright 2018 SchedMD LLC
http://www.schedmd.com

New TRes reporting options

- TresUsage{In|Out}{Ave|Min|Max|Tot}
  - NOTE: When using with Ave[RSS|VM]Size or their values in TRESUsageIn[Ave|Tot], they represent the average/total of the highest watermarks over all ranks in the step. When using sstat they represent the average/total at the moment the command was ran.
  - NOTE: TRESUsage*Min* values represent the lowest high water mark in the step.
- TresUsage{In|Out}{MinNode|MinTask|MaxNode|MaxTask}
  - Node/task that reached min/max usage

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Other, for users

- Disable local PTY output processing when using “srun --unbuffered”. Prevents “\r” insertion to output string
- Avoid terminating other processes in a task group when any task is core dumping to avoid incomplete OpenMP core files
- srun command returns the highest signal of any task
- Append “, with requeued tasks” to end of job array “end” email when any task is requeued. This is a hint to use “sacct --duplicates” to see all job accounting information

Copyright 2018 SchedMD LLC

Other, for users (continued)

- Add salloc/sbatch/srun option of --gres-flags=disable-binding to disable filtering of CPUs with respect to generic resource locality
  - This option is currently required to use more CPUs than are bound to a GRES (i.e. if a GPU is bound to the CPUs on one socket, but resources on more than one socket are required to run the job)
- Add name of partition used to output of “srun --test-only …”
  - Valuable for jobs submitted to multiple partitions
- Add job reason “ReqNodeNotAvail, reserved for maintenance”

Other, for administrators

- “scontrol reboot” enhancements
    - Add ability to specify node reason
    - Add ability to specify node state after reboot completion
    - Consider “booting” as available for backfill future scheduling and do not replace in advanced reservations
- sdiag command enhancements
    - Report outgoing message queue contents
    - Report pending job count

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Other, for administrators (continued)

- Explicitly shutdown the slurmd daemon when reboot requested
    - SlurmdParameters=shutdown_on_reboot
- Add scontrol ability to create/update TRESBillingWeights
- Calculate TRES billing values at job submission to enforce QOS DenyOnLimit configuration
- Cray: Add “CheckGhalQuiesce” to “CommunicationParameters” in slurm.conf

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Other, for administrators (continued)

- The default AuthType for slurmdbd is now "auth/munge"
- User defined triggers are now disabled by default
  - Added SlurmctldParameters option “allow_user_triggers” to enable user-defined triggers.
- SchedulerParameters’ "whole_pack" option has been renamed to "whole_hetjob" (old option still supported)

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Other, for administrators (continued)

- ConstrainKmemSpace is now disabled by default due to Linux kernel resource leaks
- cgroup_allowed_devices_file.conf no longer required
- Add acct_gather_profile/influxdb plugin to store job profiling information to InfluxDB rather than HDF5
- Added MinPrioThreshold on QOS
  - Overrides bf_min_prio_reserve slurm.conf parameter

19.05

- Completion of cons_tres plugin
- Slurm + GCP enhancements

Copyright 2018 SchedMD LLC
http://www.schedmd.com

Questions?

Copyright 2018 SchedMD LLC
http://www.schedmd.com