

# Slurm 23.02, 23.11, and Beyond

Tim Wickberg  
*Chief Technology Officer*



# Release and Development Process

# Release Cycle




- Major releases are currently made every nine months
- Version is the two digit year, two digit month:
  - 23.02, 23.11, 24.08
- Maintenance releases are made roughly monthly
  - For the most recent major release
- Releases are supported for 18 months
  - Currently: 22.05 and 23.02

# Development Process



- Most larger work is handled through sponsored projects
  - SchedMD support only covers maintenance
- Some projects - those of wider community interest - may be handled internally on a best-effort basis



# Slurm 23.02 Release

(February 2023)

# New scrun command

- Directly launch OCI-compliant container images
- Slurm's version of crun / runc
- See separate presentation from SLUG'22 for further details
  - <https://slurm.schedmd.com/publications.html>

# New --tres-per-task option



- Allow jobs to be modeled as a number of tasks, with all appropriate resource types scaled directly by the number of tasks requested
  - Task can request licenses, GRES, CPUs, memory

# License Preemption



- When running with preemption, license usage is not currently considered, and jobs will not be preempted to free up licenses
- This is an issue especially when using licenses to represent cluster-wide resources, as they won't be reclaimed to allow higher-priority work to preempt



# AllowAccounts - automatic recursion



- Update the "AllowAccounts" access control to automatically extend access to all child accounts

# Cloud nodes enhancements



- Pass list of requested features to ResumeProgram
- Reset active features on CLOUD nodes
- Allow for Node Weight to be considered on CLOUD nodes
- New flag to automatically power down "Exclusive" nodes once jobs are completed

# Reservation Enhancements



- Add a Comment field to reservations
- Show active reservations on each node in 'scontrol show node'
- Support node addition and removal from a reservation through scontrol with += and -= on the node list

# Accounting Tweaks



- New FailedNode field
  - Set for jobs that have been terminated due to a node failure
  - Help triage hardware issues

# New job completion plugin

- New jobcomp/kafka plugin

# Performance Improvements



- **Halved** the number of MUNGE interactions by slurmctld

# Flexible Node Counts



- In addition to min and max node counts, allows the user to specify acceptable node counts
  - E.g., `--nodes=20,40,80,160`
- Also allows for a step function specification
  - E.g., `--nodes=10-30:5` is equivalent to `--nodes=10,15,20,25,30`

# "Explicit" GRES Flag



- Currently, all GRES are allocated to a job when --exclusive is set
- New GRES Flag "Explicit" avoids allocating that GRES by default for --exclusive jobs
  - Will only allocate it when explicitly requested



# Debug option handling



- New 'scontrol setdebug <level> nodes=node[1-10]' sub-command
  - Allows dynamic changes to debug level on specified nodes
- 'scontrol setdebugflags flag,flag2,flag3 nodes=node[1-10]' also added

# JSON and YAML

- Greatly extended support for JSON and YAML output from user commands
- Now allows many command filtering options to be used as well

# RPC Rate Limiting

- New optional per-user RPC rate limiting mechanism
  - Backs off client commands if they're being too chatty
  - Sends new dedicated response code telling the command to sleep for a second before retrying, rather than crashing the user command
  - Can avoid having 'while true; do queue; done' overload slurmctld

# switch/hpe\_slingshot

- Developed by HPE
  - Subject to review and validation by SchedMD
- ... now works
  - 22.05 version was DOA due to libcxix library changes
- Known outstanding issues...
  - No HetJobs support at present



# Slurm 23.11 Roadmap

November 2023

# Fixing 'scontrol reconfigure'

- Plans to ensure 'scontrol reconfigure', SIGHUP, and restarting slurmctld/slurmd processes all have equivalent semantics
- Currently, certain changes cannot take effect within the process through 'scontrol reconfigure', and require a process restart
  - Which these are is undocumented, and somewhat hard to intuit
- Work to simplify these paths, and allow for additional sanity checks
- Configuration check capability expected as well

# SlurmDBD Overhaul



- The "right-left" tree data structure was used to represent the association hierarchy in a flat row-oriented fashion
  - Unfortunately, insertion and deletion is  $O(n)$ 
    - And can trigger  $O(n)$  row updates in the database
      - And  $O(n)$  updates to slurmctld
  - New "lineage" approach significantly improves performance
    - Especially when heavily scripting against external accounting systems
    - Must move slurmctld to 23.11 alongside slurmdbd to see benefits

# New auth/slurm and cred/slurm plugins



- New internal authentication and job credential plugins
  - Alternative to MUNGE
- Simple HMAC scheme (SHA-256) built off JWT
  - Separate from existing auth/jwt plugin though
  - Will require a shared secret be shared throughout the cluster
    - Similar security posture to MUNGE
- Will allow for future extension and flexibility...



# LDAP-less control plane



- Using auth/slurm, will support running the slurmctld without LDAP
  - Username, uid, gid, groups will be provided by auth/slurm

# Extended SELinux

- auth/slurm will allow for secure communication of the originating command's SELinux context

# TRES Reservations

- Allow for TRES-oriented reservations
  - E.g., reserve 200 GPUs alongside 800 CPUs

# Extensible Features



- Set of key=value pairs, with the values provided by site-specific scripts
  - Can be integers, floats, or string types
  - Values refreshed periodically (on node ping)
    - Flag can mark an extensible feature as unchanging after node boot
- Separate job submission syntax to select between these extensible features

# Relative QOS limits



- Flag allows QOS to be specified as a percentage of the cluster's total resources
  - Or an individual partition, if used as a PartitionQOS



... and Beyond

# macOS Support

- Stripped-down, but usable, Slurm install on macOS
  - Notable subsystems that macOS cannot support
    - cpu affinity
    - cgroup
    - containers
    - job\_container/tmpfs
    - ...
  - Meant mainly for testing and training



Questions?



# Next Events

- **SLUG'23**

- Hosted by Brigham Young University, in Provo, Utah, USA
- September 12-13th
- Call for Papers is out now
  - Deadline is June 9th

- **SC'23**


- Slurm Booth
- BoF - waiting on acceptance





# Audience Survey

- 
- Does anyone use the "Coordinator" setting on accounts?

- 
- How long are sites planning to continue supporting Cray Aries platforms?
    - 2023?
    - 2024?
    - 2025?
    - ... beyond?



- Container support...

- Shifter
- CharlieCloud
- podman
- Singularity / Apptainer
- Docker
- ... other?

**SCHEDMD**

The Slurm Company