Slurm 22.05, 23.02, and Beyond

Tim Wickberg
SchedMD

https://schedmd.com

Slurm User Group Meeting 2022

https://schedmd.com

Agenda - US Mountain Time (UTC-6)

Time	Speaker	Title
9:00 - 9:50	Jason Booth	Field Notes 6: From The Frontlines of Slurm Support
10:00 - 10:20	Ole Nielsen (DTU)	Pathfinding into the clouds
10:30 - 10:55	Nate Rini	OCI Containers, and scrun
11:00 - 11:20	Wei Feinstein (LBNL)	LBNL Site Report
11:30 - 11:55	Nick Ihli	Cloudy, With A Chance of Dynamic Nodes
12:00 - 12:20	Kota Tsuyuzaki (NTT)	Burst Buffer Lua Plugin for Lustre
⇨ 12:30 - 12:55	Tim Wickberg	Slurm 22.05, 23.02, and Beyond

Welcome

- Seven separate presentations, seven separate streams
- 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 staff
  - Tim McMullan, Ben Roberts, and Tim Wickberg
  - Also identified by the little wrench symbol next to their name

Slurm 22.05, 23.02, and Beyond

Tim Wickberg
SchedMD

Slurm 22.05 Release

https://schedmd.com

"Preferred" node constraints

- A list of optional ("soft") constraints to be considered when selecting nodes for a job
  - New "--prefer" option to salloc/sbatch/srun
  - Job launch will prefer those nodes, if possible to satisfy immediately
  - Traditional "hard" constraints (--constraint) will always be respected

cgroup v2 support

- Added support for cgroup v2
  - Only cgroup v1 was supported in 21.08 and older
  - Will auto-detect cgroup v1 or v2 support on the system
    - and default to v2 if available
- A number of distributions have moved to deprecate or disable v1 support, so sites are encourage to start migrating soon

Backfill for Licenses

- Licenses were previously ignored in the backfill scheduler
- By default, if licenses are currently unavailable for a job, no future reservation will be made for it
- This is obviously not ideal for sites with heavy license usage, and can lead to starvation of larger license-dependent jobs

Backfill for Licenses

- New SchedulerParameters=bf_licenses option enables license tracking in the backfill scheduler
  - Currently disabled by default, may be enabled by default in a future release

GPU Sharding

- Allow for cooperative GPU sharing between separate jobs
- Allows administrators to define a number of "Slices" for a GPU
  - Jobs can request between zero and all slices
  - All slices allocated to the job from a single GPU, cannot span between cards
- Caveat: no hardware enforcement
  - Jobs must cooperate effectively

AcctGatherInterconnect/sysfs

- Add support for gathering network statistics from OmniPath, Slingshot, and other interconnects
  - Simplified this to a single "sysfs" plugin reading stats from /sys/class/net/<interface>/statistics/
  - Able to read and aggregate stats from multiple interfaces, but will consolidate into a single ic/sysfs TRES.

Changes to LLN Support

- LLN ("Least-Loaded Node") previously defined the least-loaded nodes as those with the most idle cores
- This can lead to counter-intuitive behavior in partitions with mixed hardware
- Definition will change to LLN being the lowest proportion of allocated cores to total cores within the node

Accounting - Without Defaults

- Adding a new option to SlurmDBD to allow operation without DefaultAccounts set for every user
- Not recommended for most sites, but can simplify integration scripting with external accounting systems

slurmscriptd enhancements

- Move MailProg handling into slurmscriptd
    - Significantly improves slurmctld performance on high-throughput systems

Batch / Env Storage

- Split into a separate table
- Store hash of batch script / environment, and de-duplicate based on the hash
- Store last-used timestamp in the table
    - Allows for future Purge options to clean up

Truly Dynamic Nodes

- Move away from current FUTURE node handling
  - Support truly dynamic node addition and removal from the cluster
- Allow for better integration with, e.g., cloud systems where nodes are ephemeral

REST API

- A number of minor changes and bug fixes
- See https://slurm.schedmd.com/openapi_release_notes.html

Slurm 23.02 Roadmap

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

scrun

- Additional native container capabilities
- See Nate's presentation from earlier for further details

slurmscriptd

- Continue moving all external hooks from slurmctld into slurmscriptd
    - Massive performance benefits for large-scale and high-throughput environments from reduced fork()+exec() overhead

AllowAccounts - automatic recursion

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

… and Beyond

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

Questions?

Thank You!

- Thank you to all the presenters!
  - Especially to the community presenters
- Slides will be on the Slurm Publication Archive shortly
  - https://slurm.schedmd.com/publications.html

Next Events

- SC22, Slurm Booth - 1043

SC22

Dallas, TX | hpc accelerates.

Next Events

- SLUG'23 will be **in person**, in September 2023
    - And we'll avoid conflicting with NVIDIA GTC
- Look for announcements and call for papers on the slurm-user and slurm-announce mailing lists in the spring

End of Stream

https://schedmd.com