Slurm Community Birds-of-a-Feather SC'25

Danny Auble and Tim Wickberg
SchedMD LLC

slurm | SCHEDMD

Agenda

- Welcome
- Survey Introduction
- Slinky v1.0.0
- Slurm 25.05
- Slurm 25.11
- ... and Beyond
- Survey Results and Discussion
- Open Forum

Slides

- Slides from today - and from the booth talks - will be posted online shortly:
    - https://www.schedmd.com/publications/
- The BoFs are also being livestreamed and recorded by SC
    - Danny will be monitoring the digital experience and relaying any questions
- A version of this roadmap will be available on the SchedMD YouTube channel in early December
    - https://www.youtube.com/SchedmdSlurm

Questions?

- Feel free to ask throughout
- But - please use one of the microphones
  - Allows the folks in the room, as well as folks online, to hear your questions
- There is a panel starting in this room at 1:30pm
  - We're happy to continue discussion out in the hallway

Community Survey

Slurm Community Survey - SC'25

- https://schedmd.com/survey
- Results will be discussed towards the end of the BoF

Slinky v1.0.0 is now available!

Slinky v1.0.0

- Completion of initial design for the Slurm Operator and Slurm Bridge
- Very brief summary:
  - Slurm Operator is SchedMD's reference deployment of Slurm within Kubernetes
  - Slurm Bridge bring Slurm's scheduling wherewithal to the Kubernetes scheduling API, and allows for converged systems running both Kubernetes and Slurm workloads
- Separate documentation site: slinky.schedmd.com
  - Generated automatically through ReadTheDocs from the repo, feel free to submit PRs to update or correct
- Alpha version of the CPU DRA driver is available now
  - Allows for HPC-style CPU affinity management for Kubernetes workloads
  - https://github.com/kubernetes-sigs/dra-driver-cpu

Slurm 25.05, 25.11, and Beyond

Tim Wickberg
Chief Technology Officer

Development Cycle

Release Cycle

- Major releases are made every six months
- Version is the two digit year, two digit month:
  - 25.05 - May 2025
  - 25.11 - November 2025
  - 26.05 - May 2026
- Major releases are supported for 18 months
  - Currently: 25.11, 25.05, 24.11
  - Maintenance releases come out monthly, give or take
    - Usually only for the most recent major release
  - Security releases are made for all supported major releases

Slurm 25.05 - May 2025

Network Traffic Encryption

- Encrypt all Slurm traffic
  - Optional
  - Uses AWS's s2n-tls project as an interface to OpenSSL or other crypto suites
    - FIPS mode available if desired
- New "certmgr" plugin interface can be used for node certificate management and signing

slurmrestd - HTTPS support

- Directly provide https support through slurmrestd
- Orthogonal to encrypting Slurm's internal communication

Multiple Network Topologies

- New configuration file allows for the definition of "named" network topologies
- These named network topologies can be configured per-partition
    - Each topology can be used for one-or-more partitions

Multiple Network Topologies - Example topology.yaml config


```yaml
- topology: topo1
  cluster_default: True
  tree:
    switches:
    - switch: sw_root
      nodes: node[01-04]
- topology: topo2
  cluster_default: False
  block:
    block_sizes:
      - 4
      - 16
    blocks:
    - block: b1
      nodes: node[01-04]
    - block: b2
      nodes: node[05-08]
    - block: b3
      nodes: node[09-12]
    - block: b4
      nodes: node[13-16]
- topology: topo3
  cluster_default: False
  flat: True
```

Dynamic Network Topology

- For cloud or dynamic nodes
- New slurmd option allows the network location to be specified
  - Directly-adjacent switch must be provided, as well as any intermediate switch levels that are not part of the static topology configuration
  - If using multiple topologies, each must be enumerated
  - `slurmd -Z –conf="topology=topo1:sw_root"`

Changes for the topology/block Plugin

- Allow gaps in the block topology definition
- Capture the --segment size in accounting
  - Show in 'scontrol show job'
  - Add SLURM_JOB_SEGMENT_SIZE envvar to batch step

Hierarchical Resources

- New resource type
  - Akin to licenses (and uses some of the license subsystem), but tracked alongside a node hierarchy
    - Hierarchy may map to network topology, but is independent
    - Configuration syntax shared with multiple topology syntax
- Two current modes supported
  - Mode_1
    - Resources need to be available within one of the layers the node(s) belong to
    - E.g., job needs a single switch to provide hardware offload support
  - Mode_2
    - Equivalent resources must be available from all layers the node(s) belong to
    - E.g., job needs equal service from all network switches involved in the job

Expansion of jobcomp/kafka plugin

- New "jobstart" capability
  - Same data as would be sent at job completion
    - But now sent at job start time as well
  - JobCompParams of 'enable_job_start' and 'topic_job_start=<topic>'

PMIx MPMD Support

- Translate heterogeneous step launch into PMIx MPMD structures

Support for OR operator in license requests

- Flexibility in license requirements through OR operator support between different suitable licenses

Task Management

- New srun --wait-for-children option
  - Allows process to fork() off child processes then exit
    - Default behavior is to consider the step complete at this point
      - And terminate any additional child processes
    - Instead, step will not be marked as complete until all child processes exit
  - Requires use of cgroup/v2 plugin

HPE Slingshot Improvements

- Support network setup with over 252-tasks-per-node
  - Requires updated libcxi
- Support mTLS authentication for fabric management
  - Fabric management is used for collectives support

Reservations

- Support for creating Reservations through REST API
- Capture job --reservation request in accounting data

External Orchestrator Integration

- Support for "external" jobs
  - No batch script associated with the job
  - Will not trigger Prolog/Epilog scripts
- Support for "external" nodes
  - Created through REST API or "scontrol create NodeName=..."
  - No corresponding slurmd process
    - State must be externally managed
  - Can only be used by "external" jobs

Slurm 25.11 - November 2025

Expedited Requeue

- Allow jobs to requeue faster - *expeditiously*
- Designed to allow jobs with failed hardware to restart almost immediately
  - Avoid having healthy allocated nodes re-assigned to other work
- By default, Slurm will delay requeued jobs for 2 minutes before considering them again
  - Delay is cred_expire (default 120 seconds), can be longer on some systems
  - Required moving underlying job management from JobID to SLUID

Expedited Requeue

- Expedited requeued job is treated as having "infinite" priority
- `sbatch --requeue=expedite my_job.sh`
  - Needs SchedulerParameters=enable_expedited_requeue
- Expedited requeue triggered by:
  - Node failure
  - Batch job returning non-zero, and one or more Epilogs returning non-zero
    - Used to distinguish application failure versus hardware failure

OpenMetrics (Prometheus) support

- OpenMetrics telemetry now directly available from slurmctld
  - Accessible from SlurmctldPort
    - No need for a separate TCP port
- Similar data to what the Slurm Exporter part of the SlinkyProject was publishing
  - Removes need for data to be send and repackaged from slurmctld -> slurmrestd -> the Slurm Exporter
- On by default, can be disabled with CommunicationParameters=disable_http

New namespace plugin interface

- Replace job_container/tmpfs with namespace/tmpfs plugin
    - Still uses job_container.conf for configuration
- New namespace/linux plugin adds support for managing pid and user namespace in addition to filesystem
    - New namespace.yaml configuration file format

Reservation Access Control

- Add AllowQOS and AllowPartition access controls to reservations
  - Especially useful alongside "magnetic" reservations

Retroactive changes to energy accounting data

- `sacctmgr modify job jobid=12345 set alloctres=energy=1234`
- Allow external energy accounting data to be persisted alongside Slurm's accounting records

Support += and -= syntax when modifying GrpTRESMins

- `sacctmgr update user bbaggins account=root set grptresmins+=billing=10`

Hierarchical Resources - Mode 3

- For mode 3, resource usage is summed into all layers of the hierarchy
- Can be used for power-capping
- Additional syntactic sugar added to make mapping hardware into the topology easier:
  - A new "base" option can be set on each layer
    - Automatically subtracted from available resources
  - Allow non-node objects to be defined on the layer with corresponding usage
    - E.g., fixed power demands per rack, network resources reserved for non-compute usage
  - Allow definition of "variables" to map arbitrary names to administratively-defined resource counts
    - E.g., "maxp: 1000", allowing users to specify --resources=power:maxp
    - Variables could be modified between different clusters, increasing portability of user job scripts
- See https://slurm.schedmd.com/resources.yaml.html for config examples

Block Topology - Segment Management

- Add --spread-segments option
  - Ensure segments are in separate blocks, rather than opportunistically packed together
- Add --consolidate-segments option
  - Ensure segments are grouped within the minimum number of higher-level blocks
  - Limit resulting system fragmentation

Block Topology - Nodelist options

- Add limited support for --nodelist and --exclude options on the job allocation

switch/nvidia_imex

- Allow jobs to request independent IMEX channels for each segment
  - `--network=unique-channel-per-segment`
- The job IMEX channel is now set for batch and interactive steps
  - Previously only set for multi-node steps, now always configured

Configless Updates

- Added support for sending TaskProlog/TaskEpilog

Conmgr - Asynchronous Communication

- SlurmctldParameters=enable_async_reply
  - Improves concurrency and RPC throughput
  - Threads do not block waiting for responses to be sent to the client, instead they return to the worker pool immediately
  - Culmination of two years of work into refactoring Slurm's RPC subsystems into the new "conmgr" mode
  - Not on by default in 25.11, but will be in the future

… and Beyond

"minibatch" operation

- Extend the new stepmgr code to allow steps to queue
  - Stepmgr would dispatch the command or script to launch directly
    - Rather than the srun command

```
$ srun --batch my_step_script.sh
Submitted batch step 5609.0
```

SLUID - Slurm Lexicographically-sortable Unique ID

- 64-bit identifier for each job
  - Replaces db_index
    - Available on systems without SlurmDBD
  - Changes on each requeue
    - JobID does not - this is why the tuple of (JobID, Start Time) is what needs to be tracked externally
- Example: s8FXJCCS3F9Z00
  - Leads "s", then 13 base-32 characters (0-9, A-Z excluding ILOU)
  - "sacct --format=sluid" is the one place you can see these directly in 24.11

Removals

- Expect to completely replace all proprietary GPU plugins with ones built on devfs/sysfs
  - Finish replacement of gpu/nvml with gpu/nvidia
    - Working with NVIDIA to expose device-to-device links, usage data
  - Remove gpu/rsmi
    - Starting discussions with AMD on potential replacement with new gpu/amd plugin
- SlurmctldParameters=enable_rpc_queue option is deprecated
  - SchedMD's internal benchmarks indicates that option is detrimental to system performance starting in 25.05, and we expect to remove this entirely in a future release

Binary packages

- May offer official RPM and DEB packages in the future
  - Thus survey questions on distribution usages
- Already supporting container images for SlinkyProject
  - Packaging would simplify container image generation

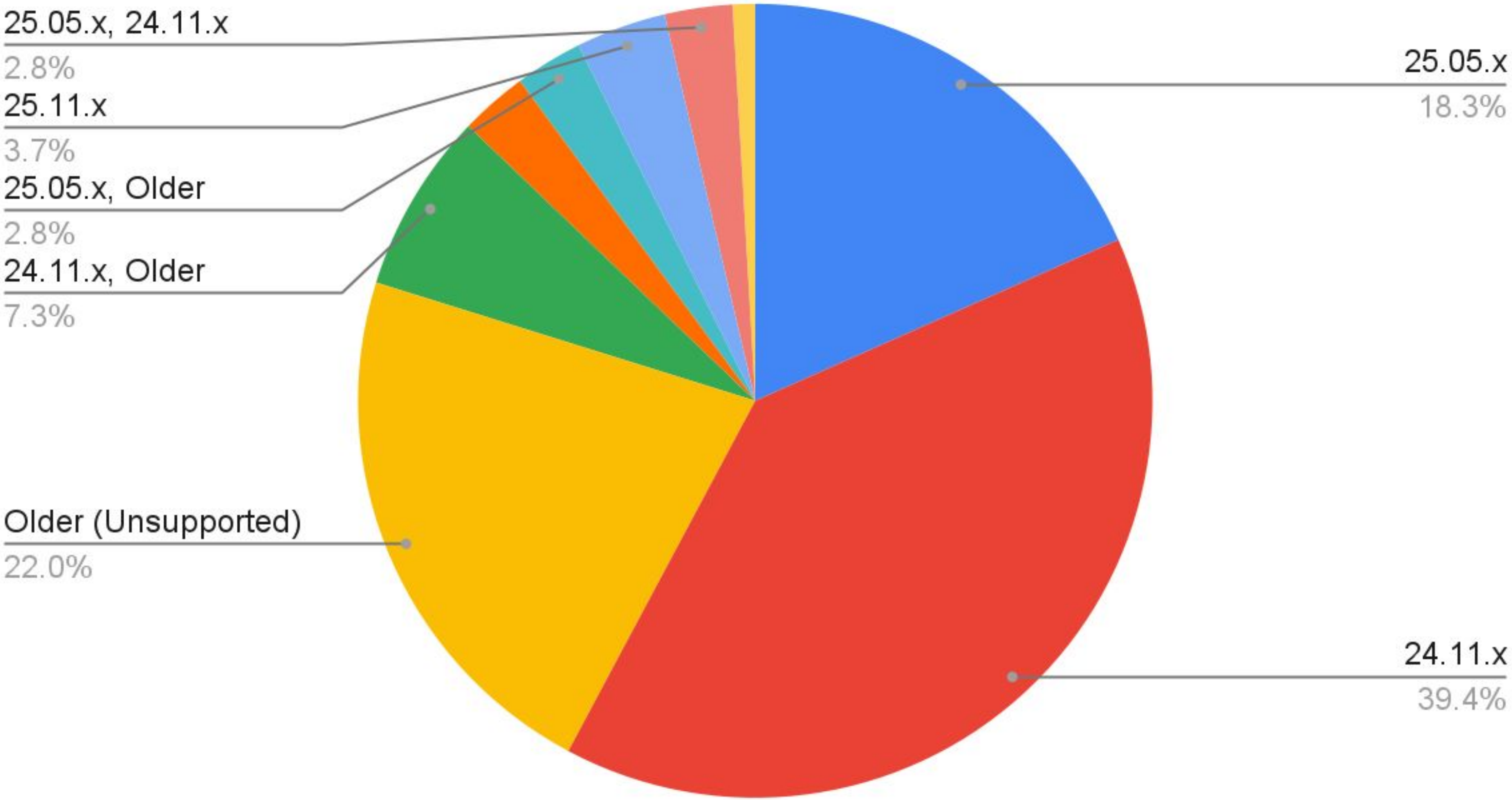
Survey Results

Most concerning response...

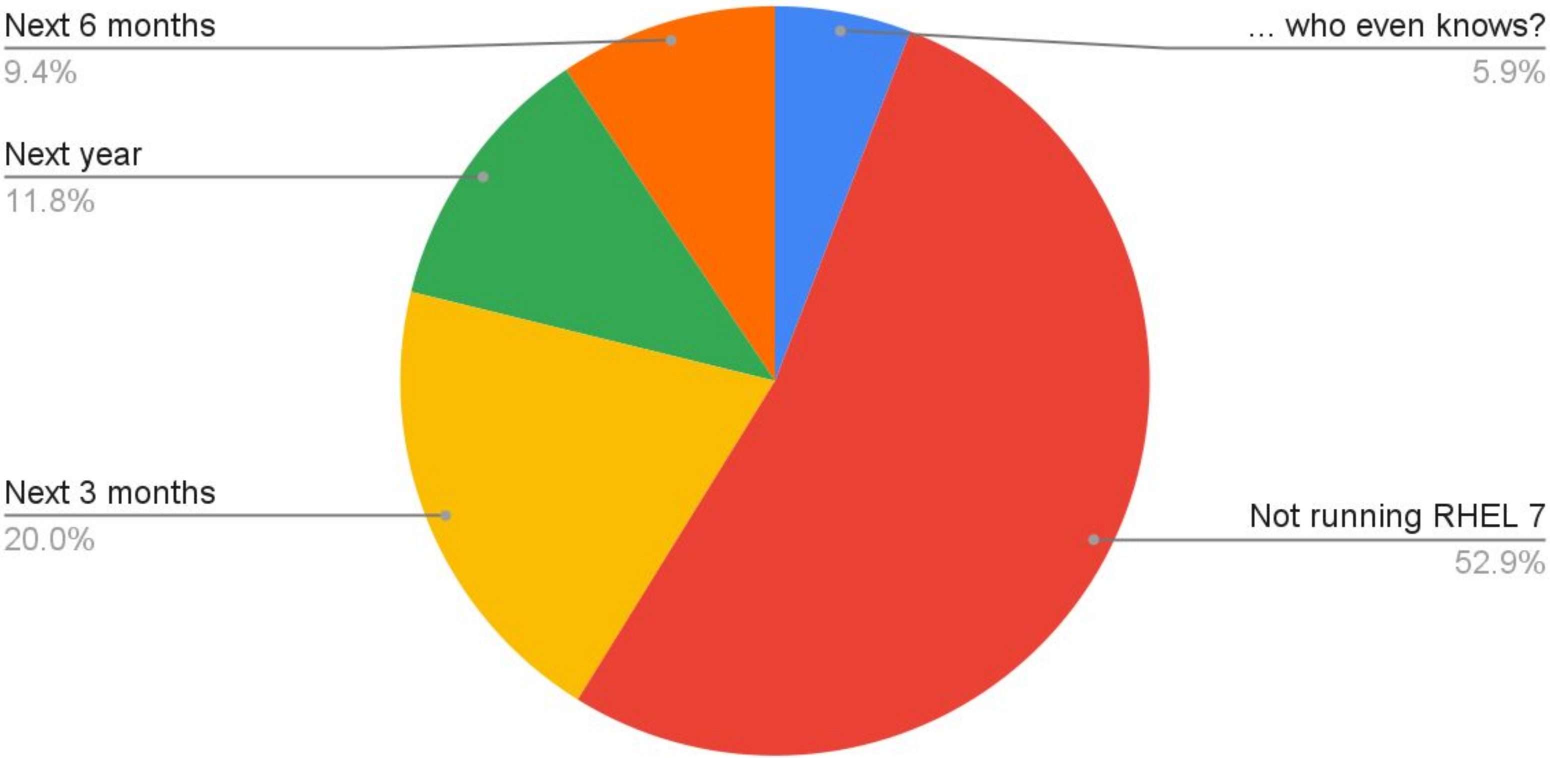
Survey asked for output from "`scontrol show config`"...

```
Slurmctld(primary) at redacted1 is DOWN
Slurmctld(backup) at redacted2 is UP
***************************************
** RESTORE SLURMCTLD DAEMON TO SERVICE **
***************************************
```

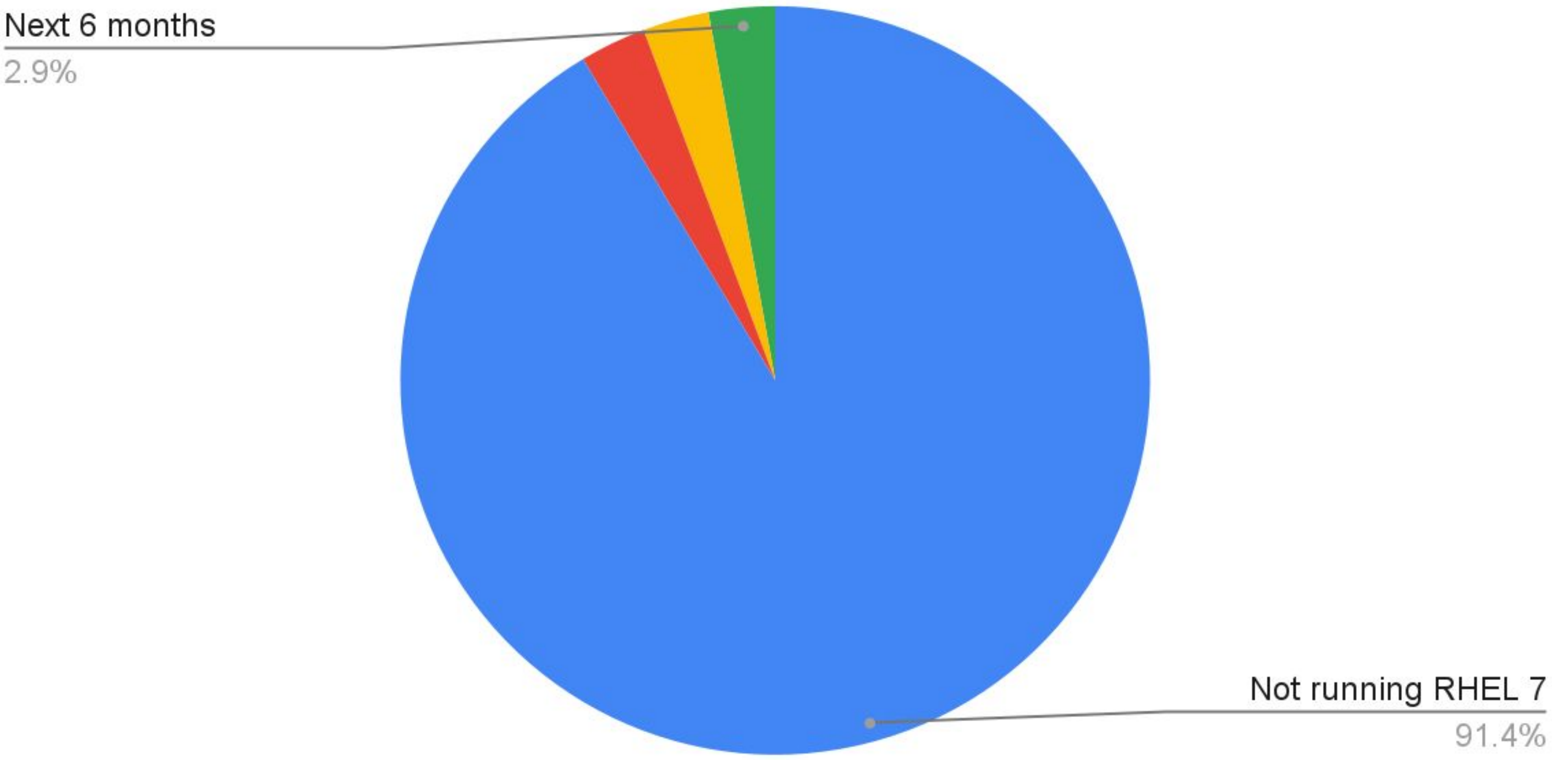
Which Slurm release(s) are you currently running in production?



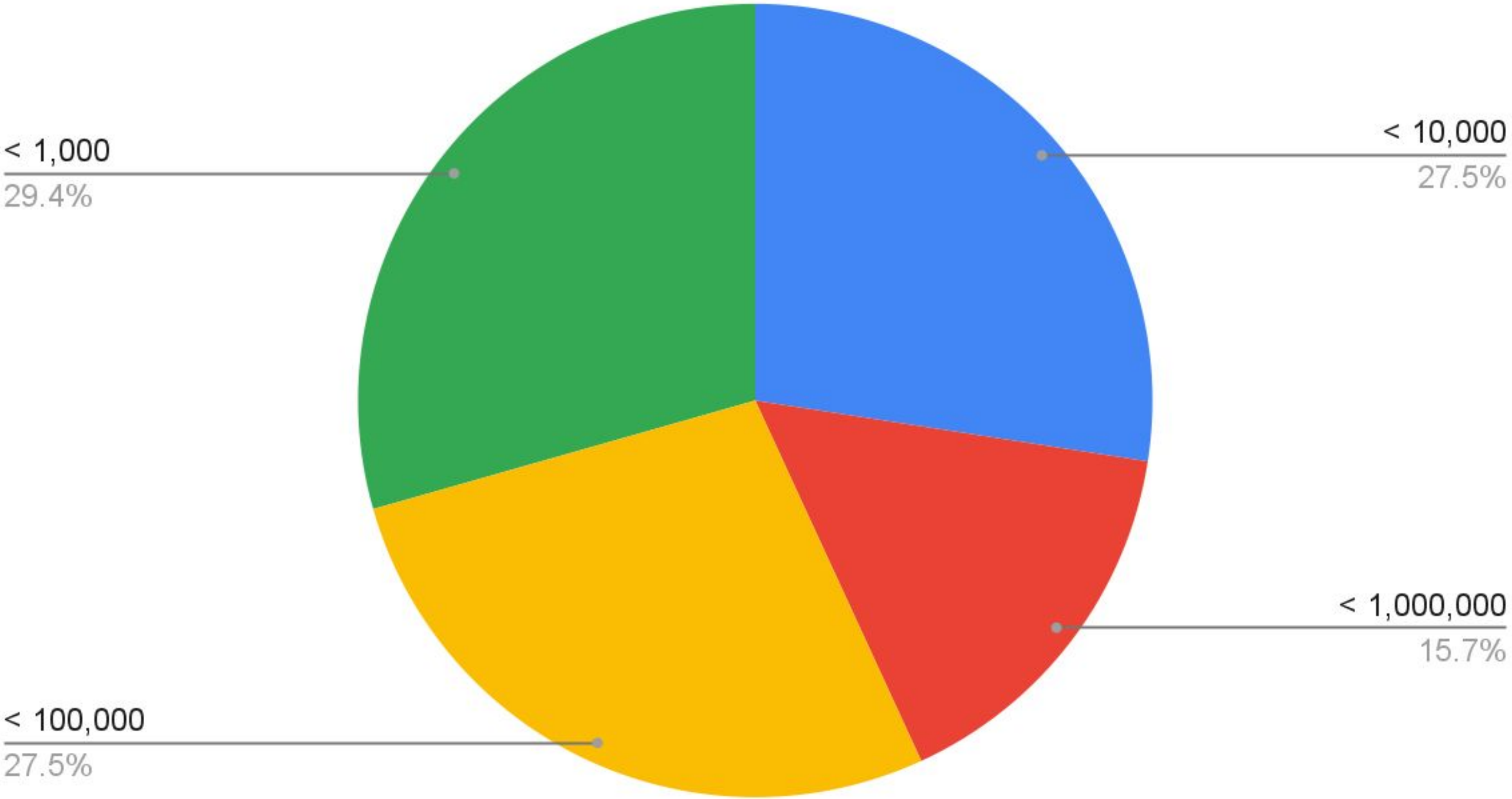
SC24 - If you're still running RHEL 7 (or derivatives), how long until you expect to migrate to something modern?



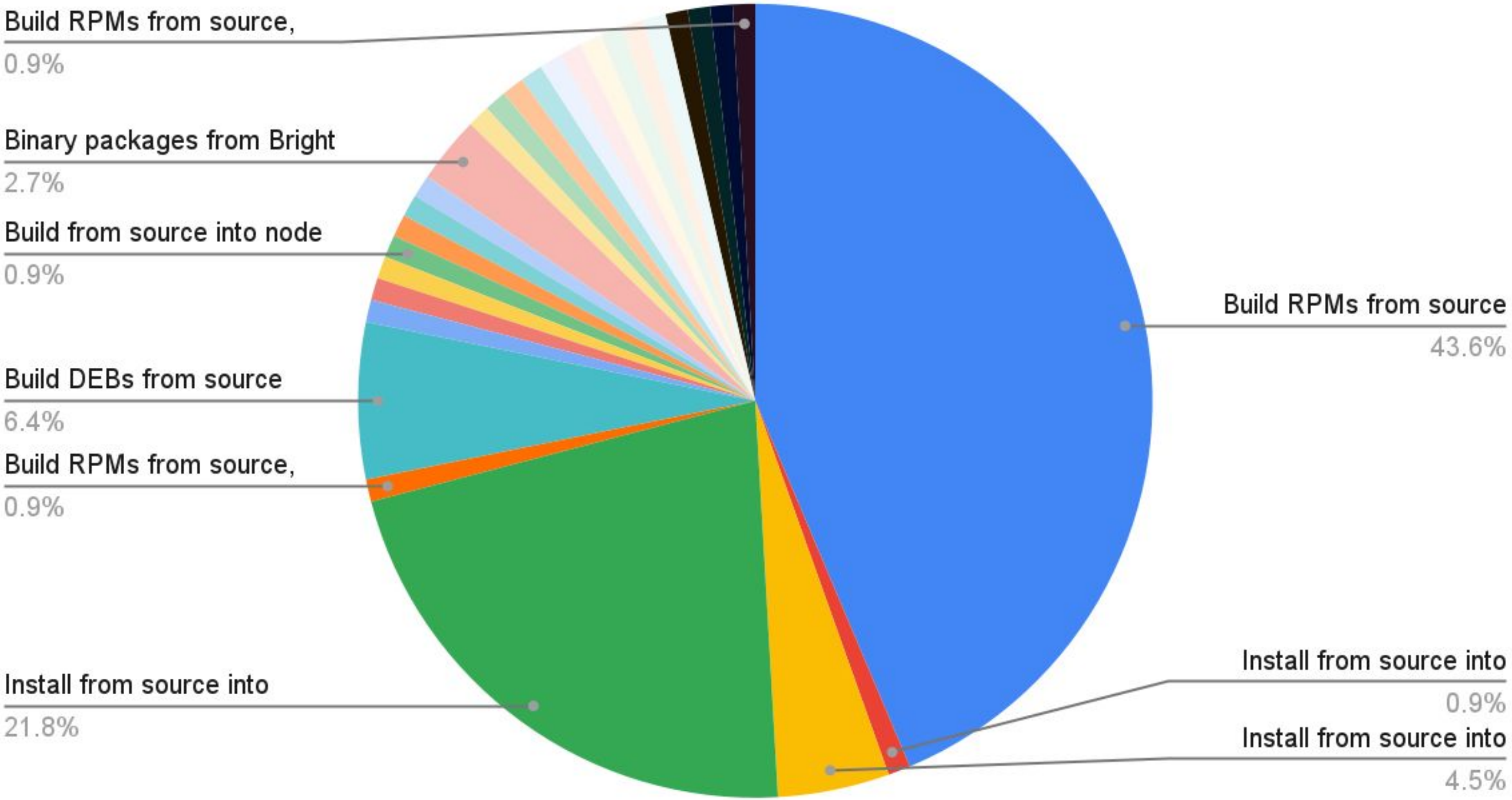
If you're still running RHEL 7 (or derivatives), how long until you expect to migrate to something modern?



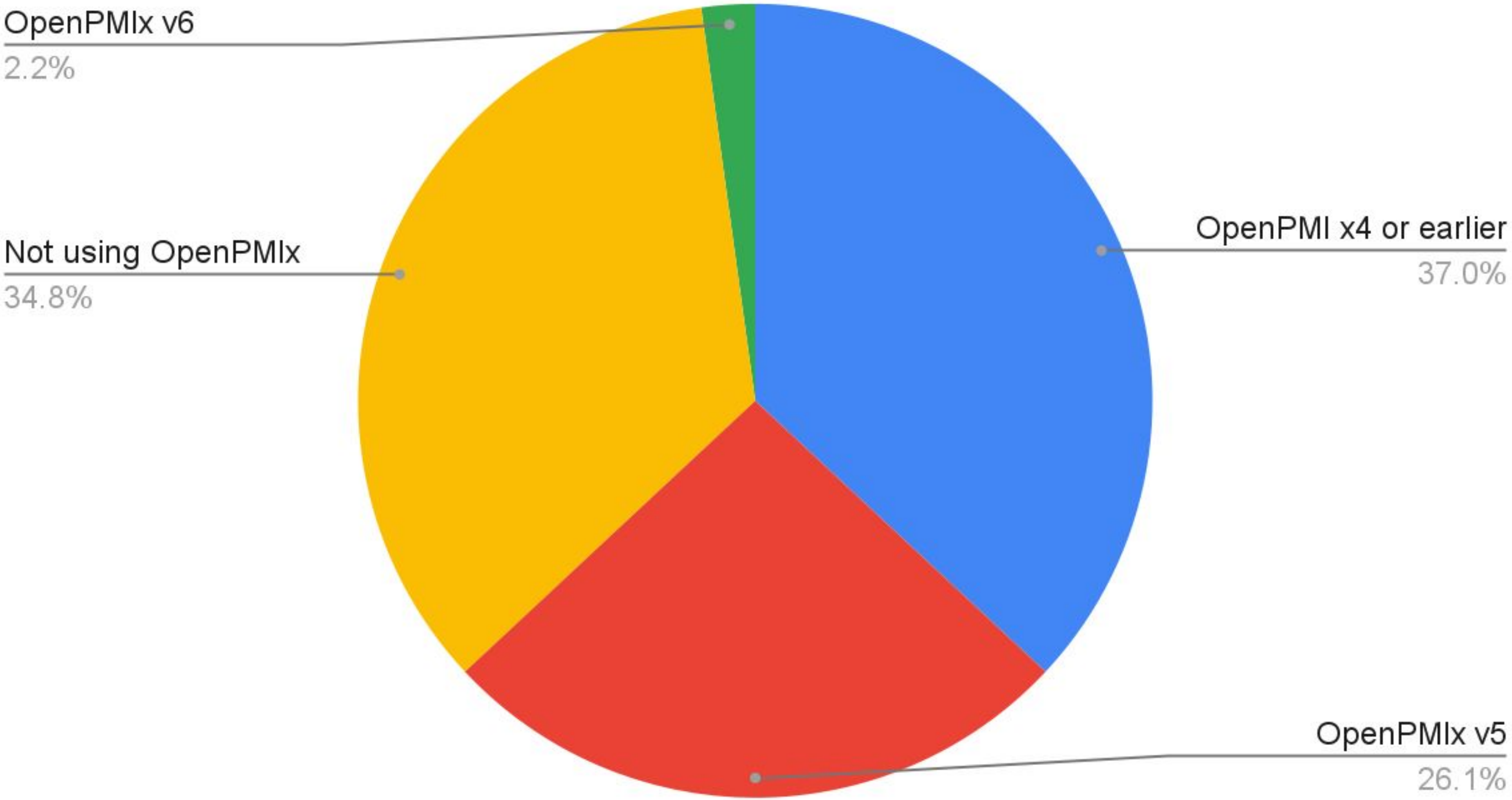
Roughly how many jobs per day does your primary system run?



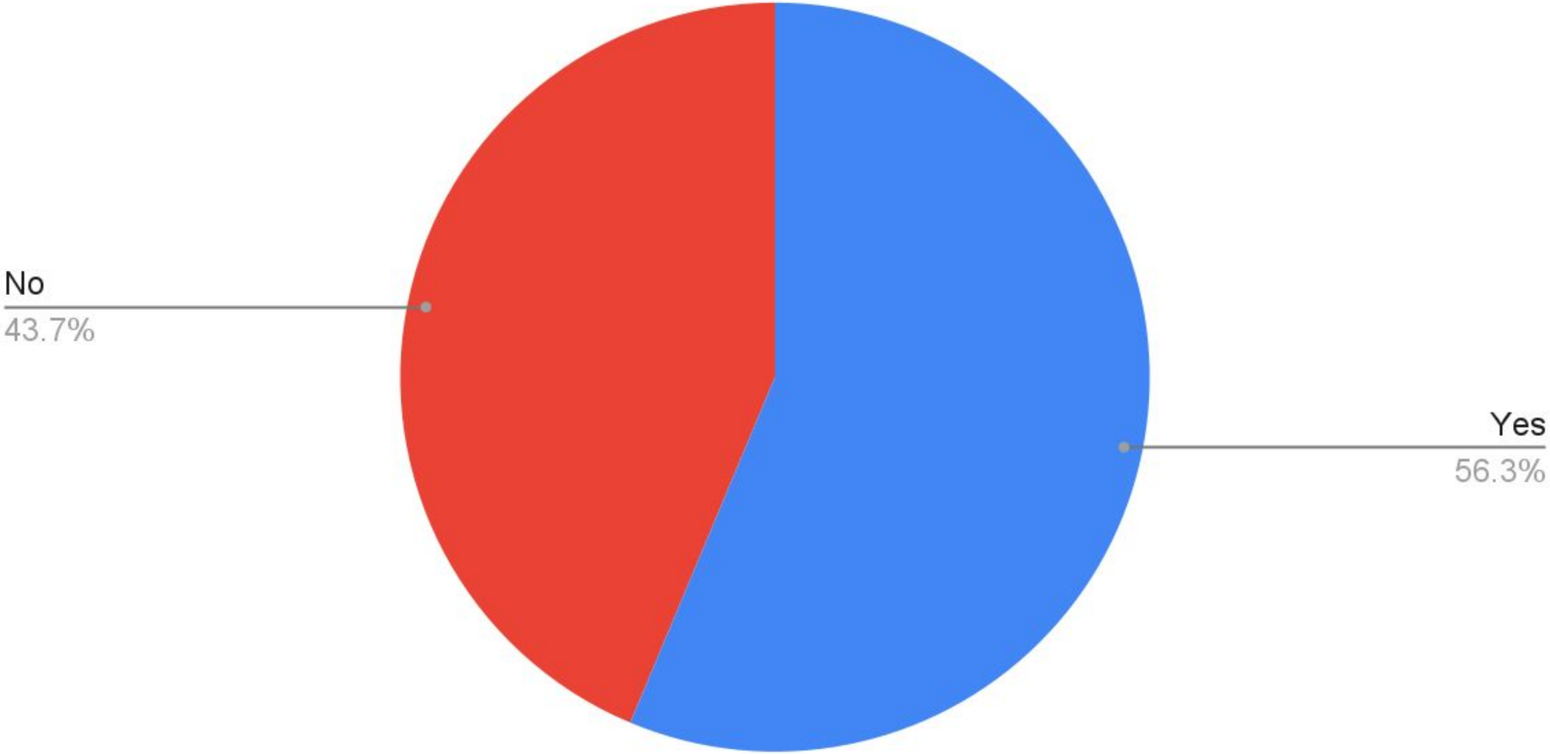
How do you manage your Slurm installation?



Count of Which OpenPMIx release are you using?



Do you have PrologFlags=x11, contain, and/or alloc set in your configuration?



Open Forum

SCHEDMD
The Slurm Company