REST API
and also
Containers

Nate Rini
SchedMD

Slurm User Group Meeting 2021

Agenda

All times are US Mountain Daylight (UTC-6)

Time	Speaker	Title
9:00 - 9:50	Jason Booth	Field Notes 5: From The Frontlines of Slurm Support
10:00 - 10:25	Nate Rini	REST API *and also* Containers
10:30 - 10:50	Marshall Garey	burst_buffer/lua and slurmscriptd
11:00 - 11:25	Nick Ihli	Slurm in the Clouds
11:30 - 11:50	Tim Wickberg	Slurm 21.08 and Beyond

Welcome

- Five separate presentations, five separate streams
- Presentations will remain available for at least two weeks after SLUG'21 concludes
- 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
- Questions will be relayed to the presenter by the moderators
  - Some may be deferred to the end if they cannot be relayed in a timely fashion
  - Or some may be answered by the moderators in chat directly

REST API
and also
Containers

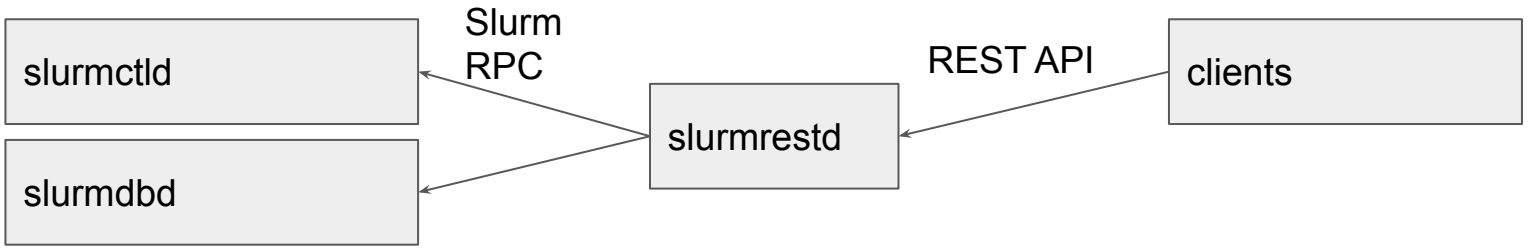
Nate Rini
SchedMD

Contents

- What is REST API & OpenAPI
- slurmrestd in Slurm 20.11
- slurmrestd in Slurm 21.08
- JSON/YAML from sacct, squeue, sinfo
- JWT Authentication
- Containers in 21.08
- Scaleout

What is the Slurm REST API

A tool that will translate JSON/YAML over HTTP requests into Slurm RPC requests.



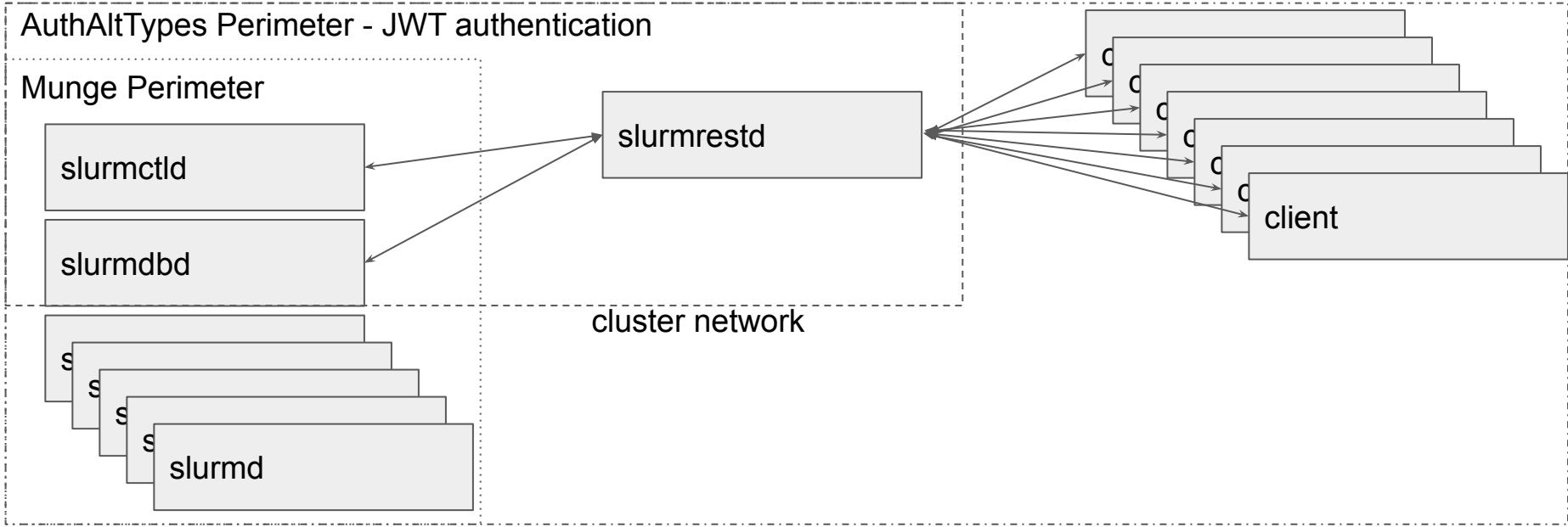
OpenAPI Compliance

OpenAPI is (arguably) becoming the industry standard for how to document a REST API. slurmrestd attempts to conform strictly to OpenAPI standards allowing 3rd party clients that can be used to generate OpenAPI clients:

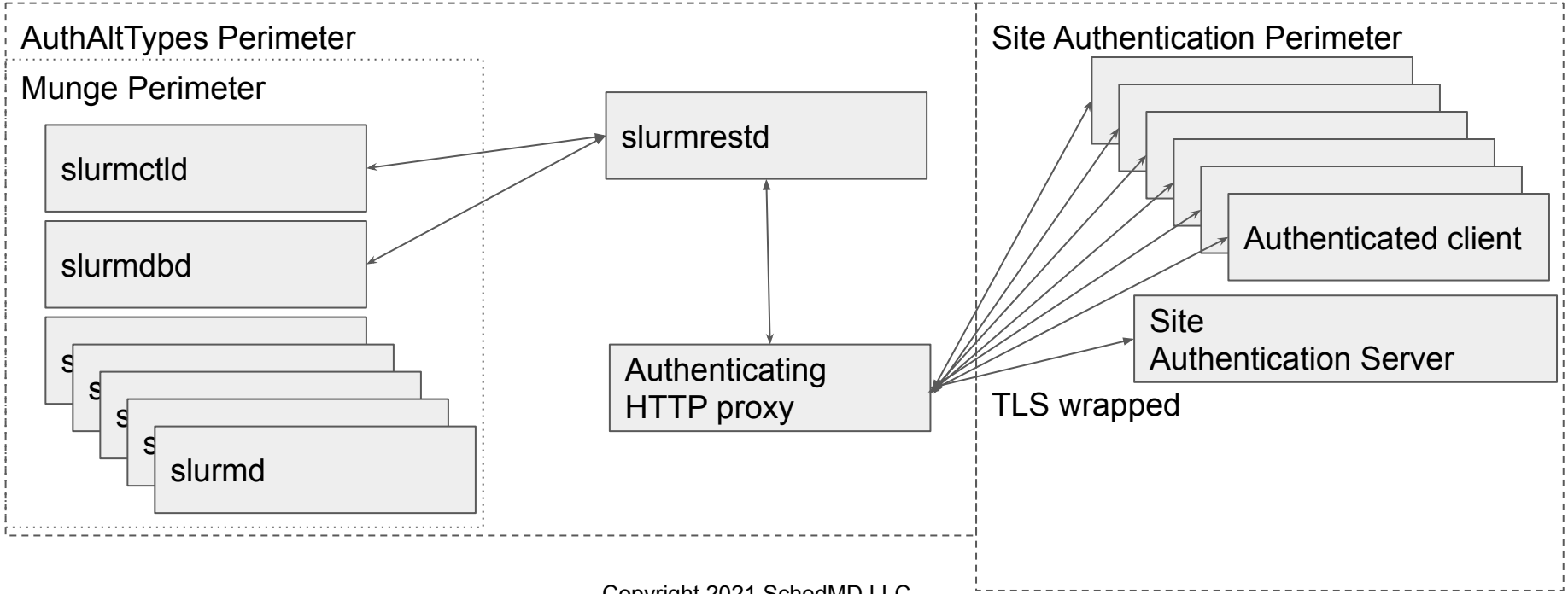
- https://openapi-generator.tech/
- https://swagger.io/tools/swagger-codegen/
- https://openapi.tools/#sdk (larger list here)

SchedMD does not provide or support any specific OpenAPI client.

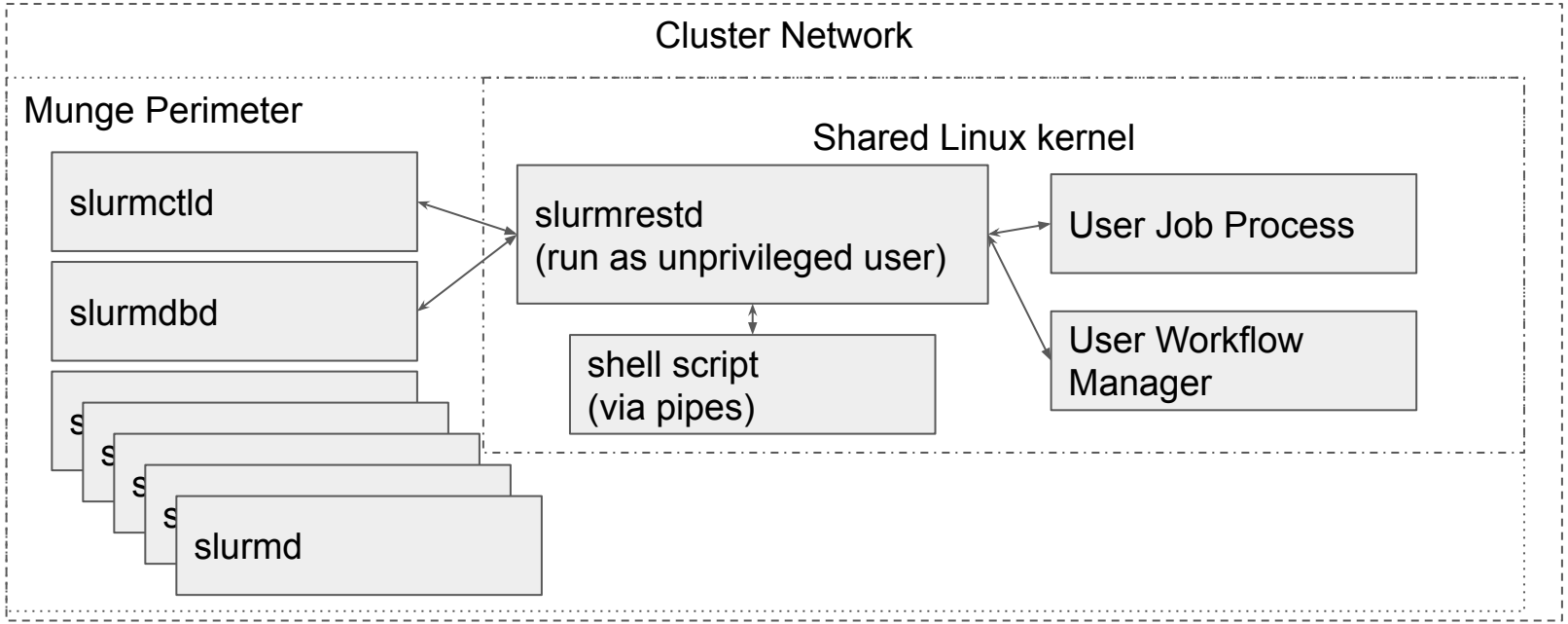
Slurm REST API Architecture (rest_auth/jwt)



Slurm REST API Architecture (rest_auth/jwt + Proxy)



Slurm REST API Architecture (rest_auth/local)



slurmrestd in Slurm 20.11

- Slurmdbd support (openapi/dbv0.0.36)
  - Ability to symmetrically update, query and load (from the same plugin version):
    - Accounts [sacctmgr show accounts]
    - Associations [sacctmgr show assoc]
    - Clusters [sacctmgr show clusters]
    - Jobs [sacct]
    - QOS [sacctmgr show qos]
    - TRES (view only - IDs are immutable) [sacctmgr show tres]
    - Users [sacctmgr show users]
    - Wckeys [sacctmgr show wckeys]
    - diag (view only) [sacctmgr show stats]
  - Data dumped from slurmrestd tends to have all information available
    as it is expected that clients can easily ignore unwanted data.
    Resultant data may be best measured in gibibytes.

slurmrestd in Slurm 20.11

- Testing against openapi-generator (v4.x)
  - While Slurm was OpenAPI standard compliant, the currently available generators were unable to implement functional clients due to their own internal limitations
  - Multiple changes were introduced to the OpenAPI specification to avoid client generation issues
  - New Slurm unit testing to verify compatibility
    - Note: v5.x of openapi-generator generates breaking changes for client API from 4.x (at least with python)
- Generated OpenAPI documentation
  - https://slurm.schedmd.com/rest_api.html
  - Include pre-generated documentation from the OpenAPI specification in Slurm's html documentation to make it easier for developers to implement clients. Previously, this had to be generated via one of the OpenAPI generators.

slurmrestd in Slurm 21.08

- New change/release notes for slurmrestd
  - https://slurm.schedmd.com/rest_release_notes.html
  - Provides list of API changes to assist client developer to update between releases
    - Changes between release (so far) have been mostly minor.
    - OpenAPI specification is tricky to diff between versions due to large number of name changes.
- Add update_time to allow clients to quickly detect changes
- Add dumping of reservations [scontrol show reservations]
- Deprecation of v0.0.35
  - The original release plugin of slurmrestd has been marked as deprecated and will be removed in Slurm 22.05 release.

JSON/YAML from sacct,squeue,sinfo

- Slurmrestd uses content (a.k.a. openapi) plugins. These plugins have been made global to allow other parts of Slurm to be able to dump JSON/YAML output.
- New output formatting (limited to these binaries only):
  - sacct --json or sacct --yaml
  - sinfo --json or squeue --yaml
  - squeue --json or squeue --yaml
- Output is always same format of latest version of slurmrestd output.
  - Formatting arguments are ignored for JSON or YAML output as it is expected that clients can easily pick and choose what they want.
- slurmrestd is not involved with this JSON/YAML output.

JWT Authentication

- Added support for JWT in slurmdbd (Slurm 20.11+)
  - Clients can now send queries to slurmdbd using JWT authentication.
  - Location of jwt_hs256.key is now configurable, as slurmdbd does not have access to StateSaveLocation.

```
[AuthAltParameters=jwt_key=$PATH_TO_KEY]
```

- Disable User JWT generation (Slurm 20.11+)
  - Sites can now disable unprivileged users generating tokens:
    - AuthAltParameters=disable_token_creation
  - Documented examples on how to generate tokens outside of Slurm.
  - Sites can now enforce rules on who and what tickets are created.
- Support for JWKS/RS256 (Slurm 21.08+)
  - Allows for authentication server to hold private key and Slurm to validate using the public key.
  - Possible to use existing external authentication systems, such as Amazon Cognito.

OCI Container Support (21.08)

Technical Preview

This functionality has been added as a technical preview as we continue to work out all the features and site requirements. RFE tickets are always welcome.

Functionality and interfaces may change dramatically between releases.

- Container is a little ambiguous of a term. This is specifically Open Container Initiative (OCI) containers which follow the set of standards here:
  - https://github.com/opencontainers
- Slurm container documentation:
  - https://slurm.schedmd.com/containers.html
- Note: job_container/tmpfs is independent from OCI Container functionality

OCI Container Support (21.08)

- Slurm now supports (limited) executing of OCI Containers via OCI runtimes
  - Relevant standards: OCI Runtime & OCI Image
  - OCI containers were originally developed by Docker but are now used in a few places including Kubernetes.
    - Docker appears to update the OCI standard on major releases and they are not always compatible changes
  - All OCI containers are started/controlled via an OCI runtime executable
    - There are several OCI runtimes, each of varying level of compliance with the standard.
  - Existing containers:
    - Limited OCI container support already exists for Singularity and charlie-cloud.
    - Sarus already has full OCI container support due to their use of runc.

OCI Container Support (21.08)

- Added ‘--container’ support to the following:
  - Srun
  - salloc
  - sbatch
- Added viewing job container to the following:
  - scontrol show jobs
  - scontrol show steps
  - sacct
    - If passed as part of the ‘--format” argument using “Container”
  - slurmdbd & slurmctld logs (too many places to list)

OCI Container Support (21.08)

- Slurm cgroups features apply to the OCI containers
  - All processes should be cleaned up even if the container anchor process dies or processes attempt to become daemons and detach from the session.
  - Resources usage can be hard limited and monitored
- Slurm is only going to support unprivileged containers in 21.08
  - Use existing kernel support for containers
  - Users can already call all of these commands directly
  - Containers must be able to function in an existing host network
- New ‘oci.conf’ in /etc/slurm/
  - If ‘oci.conf’ is not populated, the ‘--container’ request will only be recorded.
    - Environment variable SLURM_CONTAINER will always be set with value (if present).

OCI Container Support (21.08)

srun & salloc examples

```
$ srun --container=/tmp/centos grep ^NAME /etc/os-release
NAME="CentOS Linux"
$ salloc --container=/tmp/centos
salloc: Granted job allocation 24418
bash: cannot set terminal process group (-1): Inappropriate ioctl for device
bash: no job control in this shell
```

Note: containers have limited permissions and can result in some warnings

OCI Container Support (21.08)

sbatch example

```
$ sbatch --container=/tmp/centos --wrap 'grep ^NAME /etc/os-release'
Submitted batch job 24419
$ cat slurm-24419.out
NAME="CentOS Linux"
```

OCI Containers Images (21.08)

- User **must** prepare their OCI images per specifications:
  - https://github.com/opencontainers/image-spec
- Container images **must** already be visible on executing compute node.
  - Slurm does not copy or mount any images directly from job submission node.
- OCI image structure (folder structure)
  - Slurm only cares about the ‘config.json’ file in the root directory
  - config.json provides all the required information for Slurm to request a container instance
  - Slurm will produce a per step copy (in spool dir) of the config.json and provide it to the OCI runtime to allow the editing of the requests. This is required to be able to run different arguments inside of the container.
  - The OCI runtime handles all the details of mount types including overlays

Experimentation with Scaleout

- What is Scaleout
  - Take our official training course and walk through practical exercises with an experience instructor. Please email sales@schedmd.com to setup a training session.
  - Docker compose “pod” designed to replicate a working Slurm cluster
  - Simulates a full cluster (including cloud nodes) on your laptop
  - Cluster is isolated but will have outbound IPv4 NATed communications
  - Setup with suggested configuration, including slurmrestd as an example
- Download here: https://gitlab.com/SchedMD/training/docker-scale-out
  - Make sure to pick a branch: slurm-21.08 or slurm-20.11
  - Read the README
    - sysctl settings must be applied first!
    - docker-compose can be picky about versions
      - latest release of docker and docker-compose is highly recommended

Questions?

Next Session

- The next presentation is by Marshall Garey: "burst_buffer/lua and slurmscriptd"
- Starts at 10:30am Mountain Daylight Time (UTC-6)
- And is on a separate YouTube Live stream
- Please see the SchedMD Slurm YouTube channel for links

End Of Stream

- Thanks for watching!