Slurm ♥ Containers

CANOPIE-HPC - SC22

Tim Wickberg
tim@schedmd.com

OCI Container Support (21.08+)

- Slurm cgroups features apply to the OCI containers
  - All processes cleaned up even if the container anchor process dies, or try to daemonize and detach from the session
  - Resource usage can be hard limited and monitored
- Slurm only supports unprivileged containers
  - 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
- Per host configuration via ‘oci.conf’ in /etc/slurm/
  - Environment variables SLURM_CONTAINER and SLURM_CONTAINER_ID (23.02) available

Slurm OCI Container Support

- Added ‘--container’ (21.08) support to the following:
  - srun
  - salloc
  - sbatch
- Added viewing job container [bundle path] (21.08) and container-id (23.02) to the following:
  - scontrol show jobs
  - scontrol show steps
  - sacct
    - If passed as part of the ‘--format’ argument using “Container”
  - slurmd, slurmstepd, slurmdbd & slurmctld logs (too many places to list)

OCI Container Support (21.08+)

srun example

```
$ srun --container=/tmp/centos grep ^NAME /etc/os-release
NAME="CentOS Linux"
```

salloc example

```
$ salloc --container=/tmp/centos grep ^NAME /etc/os-release
salloc: Granted job allocation 65
NAME="CentOS Linux"
salloc: Relinquishing job allocation 65
```


Note: containers have limited permissions and can result in pseudo terminal 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 runtime proxy - scrun (23.02)

- scrun's goal is to make containers **boring for users**
  - Users have better things to do than learn about the intricacies of containers
  - Site administrators will have to do setup and maintenance on the configuration
- Use Slurm's existing infrastructure to run containers on compute nodes
- Automatic staging out and in of containers controlled by system administrators
  - End requirement that users manually prepare their images on compute nodes.
- Interface directly with OCI runtime clients (Docker or Podman or …)

OCI runtime proxy - scrun (23.02)

- Allow users to work with the tools they want while running work on the Slurm cluster
- scrun is a new CLI command to join srun, sbatch and salloc
  - But users are not expect to call it directly, designed to bolt in underneath existing container tooling

Rootless Docker config (23.02)

~/.config/docker/daemon.json

```
{
  "default-runtime": "slurm",
  "runtimes": {
    "slurm": {
      "path":
"/usr/local/slurm/sbin/scrun"
    }
  },
  "experimental": true,
  "iptables": false,
  "bridge": "none",
  "no-new-privileges": true,
  "rootless": true,
  "selinux-enabled": false
}
```

Podman config for scrun (23.02)

~/.config/containers/containers.conf:

```
[containers]
apparmor_profile = "unconfined"
cgroupns = "host"
cgroups = "enabled"
default_sysctls = []
label = false
netns = "host"
no_hosts = true
pidns = "host"
utsns = "host"
userns = "host"
```

```
[engine]
runtime = "slurm"
runtime_supports_nocgroups
= [ "slurm" ]
runtime_supports_json = [
"slurm" ]
remote = false

[engine.runtimes]
slurm = [
"/usr/local/slurm/sbin/scrun" ]
```

scrun via rootless Docker (23.02)

example:

```
$ export DOCKER_HOST=unix://$XDG_RUNTIME_DIR/docker.sock
$ export DOCKER_SECURITY=”--security-opt label:disable --security-opt
seccomp=unconfined  --security-opt apparmor=unconfined --net=none”
$ docker run $DOCKER_SECURITY -i ubuntu /bin/sh -c 'grep ^NAME /etc/os-release'
NAME="Ubuntu"
$ docker run $DOCKER_SECURITY -i centos /bin/sh -c 'grep ^NAME /etc/os-release'
NAME="CentOS Linux"
```

scrun via rootless Podman (23.02)

example:

```
$ podman run ubuntu /bin/sh -c 'grep ^NAME /etc/os-release'
NAME="Ubuntu"
$ podman run centos /bin/sh -c 'grep ^NAME /etc/os-release'
NAME="CentOS Linux"
$ podman run centos /bin/sh -c 'printenv SLURM_JOB_ID'
77
$ podman run centos /bin/sh -c 'printenv SLURM_JOB_ID'
78
```

Questions?

Shameless Plug:
Slurm and/or/vs Kubernetes - Slurm Booth (1043)
Tuesday 3:15pm, Wednesday 4:15pm