SLURM - Simple Linux Utility for Resource Management

Introduction

Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for large and small Linux clusters.

It provides three key functions:

- allocating exclusive and/or non-exclusive access to resources (computer nodes) to users for some duration of time so they can perform work,
- providing a framework for starting, executing, and monitoring work (typically a parallel job such as MPI) on a set of allocated nodes, and
- arbitrating contention for resources by managing a queue of pending jobs.

Installation

Controller

Controller name: slurm-ctrl

```
ssh csadmin@slurm-ctrl
sudo apt install slurm-wlm slurm-wlm-doc mailutils sview mariadb-client
mariadb-server libmariadb-dev python-dev python-mysqldb
```

Install Maria DB Server

```
apt-get install mariadb-server
systemctl start mysql
mysql -u root
create database slurm_acct_db;
create user 'slurm'@'localhost';
set password for 'slurm'@'localhost' = password('slurmdbpass');
grant usage on *.* to 'slurm'@'localhost';
grant all privileges on slurm_acct_db.* to 'slurm'@'localhost';
flush privileges;
exit
```

In the file /etc/mysql/mariadb.conf.d/50-server.cnf we should have the following setting:

```
bind-address = localhost
```

Configure munge

```
ssh csadmin@linux1
scp slurm-ctrl:/etc/munge/munge.key /etc/munge/
```

Node Authentication

First, let us configure the default options for the munge service:

/etc/default/munge:

OPTIONS="–syslog –key-file /etc/munge/munge.key"

Central Controller

The main configuration file is /etc/slurm-llnl/slurm.conf this file has to be present in the controller and all of the compute nodes and it also has to be consistent between all of them.

```
##############################
# /etc/slurm-llnl/slurm.conf
##############################
# General
ControlMachine=entry-node
AuthType=auth/munge
CacheGroups=0
CryptoType=crypto/munge
JobCheckpointDir=/var/lib/slurm-llnl/checkpoint
KillOnBadExit=01
MpiDefault=pmi2
MailProg=/usr/bin/mail
PrivateData=usage,users,accounts
ProctrackType=proctrack/cgroup
PrologFlags=Alloc,Contain
PropagateResourceLimits=NONE
RebootProgram=/sbin/reboot
ReturnToService=1
SlurmctldPidFile=/var/run/slurm-llnl/slurmctld.pid
SlurmctldPort=6817
SlurmdPidFile=/var/run/slurm-llnl/slurmd.pid
SlurmdPort=6818
SlurmdSpoolDir=/var/lib/slurm-llnl/slurmd
SlurmUser=slurm
StateSaveLocation=/var/lib/slurm-llnl/slurmctld
SwitchType=switch/none
TaskPlugin=task/cgroup

# Timers
InactiveLimit=0
```

```
KillWait=30
MinJobAge=300
SlurmctldTimeout=120
SlurmdTimeout=300
Waittime=0

# Scheduler
FastSchedule=1
SchedulerType=sched/backfill
SchedulerPort=7321
SelectType=select/cons_res
SelectTypeParameters=CR_CPU_Memory

# Preemptions
PreemptType=preempt/partition_prio
PreemptMode=REQUEUE

# Accounting
AccountingStorageType=accounting_storage/slurmdbd
AccountingStoreJobComment=YES
ClusterName=mycluster
JobAcctGatherFrequency=30
JobAcctGatherType=jobacct_gather/linux
SlurmctldDebug=3
SlurmctldLogFile=/var/log/slurm-llnl/slurmctld.log
SlurmdDebug=3
SlurmdLogFile=/var/log/slurm-llnl/slurmd.log
SlurmSchedLogFile= /var/log/slurm-llnl/slurmschd.log
SlurmSchedLogLevel=3

NodeName=compute-1 Procs=48 Sockets=4 CoresPerSocket=12 ThreadsPerCore=1
RealMemory=128000 Weight=4
NodeName=compute-2 Procs=48 Sockets=4 CoresPerSocket=12 ThreadsPerCore=1
RealMemory=254000 Weight=3
NodeName=compute-3 Procs=96 Sockets=2 CoresPerSocket=24 ThreadsPerCore=2
RealMemory=256000 Weight=3
NodeName=compute-4 Procs=96 Sockets=2 CoresPerSocket=24 ThreadsPerCore=2
RealMemory=256000 Weight=3

PartitionName=base Nodes=compute-1,compute-2,compute-3,compute-4 Default=Yes
MaxTime=72:00:00 Priority=1 State=UP
PartitionName=long Nodes=compute-1,compute-2,compute-3,compute-4 Default=No
MaxTime=UNLIMITED Priority=1 State=UP AllowGroups=long
```

```
root@controller# systemctl start slurmctld
```

Accounting Storage

After we have the slurm-llnl-slurmdbd package installed we configure it, by editing the /etc/slurm-llnl/slurmdb.conf file:

```
#######################################################################
#
# /etc/slurm-llnl/slurmdb.conf is an ASCII file which describes Slurm
# Database Daemon (SlurmDBD) configuration information.
# The contents of the file are case insensitive except for the names of
# nodes and files. Any text following a "#" in the configuration file is
# treated as a comment through the end of that line. The size of each
# line in the file is limited to 1024 characters. Changes to the
# configuration file take effect upon restart of SlurmDbd or daemon
# receipt of the SIGHUP signal unless otherwise noted.
#
# This file should be only on the computer where SlurmDBD executes and
# should only be readable by the user which executes SlurmDBD (e.g.
# "slurm"). This file should be protected from unauthorized access since
# it contains a database password.
#######################################################################
AuthType=auth/munge
AuthInfo=/var/run/munge/munge.socket.2
StorageHost=localhost
StoragePort=3306
StorageUser=slurm
StoragePass=safepassword
StorageType=accounting_storage/mysql
StorageLoc=slurm_acct_db
LogFile=/var/log/slurm-llnl/slurmdbd.log
PidFile=/var/run/slurm-llnl/slurmdbd.pid
SlurmUser=slurm
```

```
root@controller# systemctl start slurmdbd
```

Test munge

```
munge -n | unmunge | grep STATUS
STATUS:           Success (0)
munge -n | ssh slurm-ctrl unmunge | grep STATUS
STATUS:           Success (0)
```

Test Slurm

```
sinfo
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
debug*       up   infinite      1   idle linux1
```

Compute Nodes

A compute node is a machine which will receive jobs to execute, sent from the Controller, it runs the slurmd service.

Authentication

```
ssh root@slurm-ctrl
root@controller# for i in `seq 1 2`; do scp /etc/munge/munge.key linux-
${i}:/etc/munge/munge.key; done
```

```
root@compute-1# systemctl start munge
```

Run a job from slurm-ctrl

```
ssh csadmin
srun -N 1 hostname
linux1
```

https://slurm.schedmd.com/overview.html