MSLURM

SLURM management of multiple environments

1. Introduction

This document is intended as an overview to mslurm for administrators

Within an ordinary SLURM installation a single SLURM cluster is served by a SLURM master ("SLURM Control Node").

Consequently, certain daemons are running on a SLURM master computer:

- SLURM control daemon ("slurmctld"),
- Upon recommended use of a MySQL database also a SLURM data base daemon ("slurmdbd") and a MySQL server daemon ("mysqld") can run on SLURM master.

In case several SLURM clusters or several SLURM data bases are managed by a single SLURM master, an appropriate superstructure is necessary.

2. Idea

Such a superstructure for the management of multiple SLURM environments is done with MSLURM. Thereby several SLURM clusters - even across multiple SLURM databases - can run parallel on a SLURM master and can be administered in an easy and elegantly manner.

- MSLURM scripts are only used on the single merged SLURM master.
- The configuration of the compute or login nodes doesn't have to be changed.
- MSLURM exploits the fact that a single SLURM environment can be configured in such a way that no conflicts arise with different SLURM environments.
    - Defining different ports in the configuration files ("slurm.conf","slurmdbd.conf")
    - Creating new home directories for each cluster configuration.

3. Implementation

The script "mslurm" ("/usr/sbin/mslurm" or even "/usr/sbin/mslurmdbd"), enables both the individual and multiple SLURM daemons management and the execution of slurm commands.

To achieve this MSLURM uses the environment variable SLURM_CONF to specify the location of each slurm.conf file to each cluster.

The mslurm.conf ("/etc/slurm/mslurm.conf") file contains set up information for all the different clusters and databases, which will run on the MSLURM master.

Furthermore, suitably modified SLURM startup scripts ("/etc/init.d/slurm, /etc/init.d/slurmdbd") belong to MSLURM, as long as or because SLURM supports the variable SLURM_CONF as described in the SLURM documentation with its own startup scripts.

These SLURM startup scripts are exclusively used by the MSLURM startup scripts ("/etc/init.d/mslurm /etc/init.d/mslurmdbd" and shorter "/usr/sbin/rcmslurm /etc/init.d/rcslurmdbd"), which manage the one MSLURM environment as a whole, i.e. all defined objects.

4. MySQL

MySQL supports the use of independent parallel running MySQL databases ("mysqld_multi").

The MySQL databases used one-on-one by the SLURM databases, which are managed by an ordinary multi-managed MySQL installation ("/etc/mysql/mysql.conf").

5. Internal Dependencies

When managing a MSLURM environment consider the existing internal logical dependencies when booting

[slurmd <-] *) slurmctld <- slurmdbd <- mysqld_multi

as given by the existing installed boot scripts ("/etc/init.d/{mysqld_multi,slurbdbd,slurm}".

**) Also a MSLURM master can function as normal SLURM node.*

6. Multiple Configurations

Each of the SLURM environments of a MSLURM environment are configured just like ordinary SLURM environments except for different ports and directories.

MLSURM supports simple strings for the distinction by name between SLURM environments.

To distinguish the configuration files ("slurm.conf, slurmdbd. conf") of every SLURM environment, the files associated with the different configurations have to be placed in different directories (which are cluster name dependent).

7. Configuration Example

An individual MSLURM directory structure for the SLURM clusters "foo" and "bar", that share the database "baz", and the SLURM clusters "cluster1b", "cluster2b" and "cluster3b" the database "unionb" could then be obtained with the following SLURM configuration files:

/etc/slurm/db=baz/slurmdbd.conf:

..

LogFile=/var/log/slurmdbd-baz.log

PidFile=/var/run/slurmdbd-baz.pid

DbdPort=9119

..

/etc/slurm/cluster=foo/slurm.conf:

..

StateSaveLocation=/var/slurm-foo

SlurmdSpoolDir=/var/slurmd-foo

SlurmctldPidFile=/var/run/slurmctld-foo.pid

SlurmdPidFile=/var/run/slurmd-foo.pid

SlurmctldLogFile=/var/log/slurmctld-foo.log

SlurmdLogFile=/var/log/slurmd-foo.log

SlurmctldPort=9117

SlurmdPort=9118

AccountingStoragePort=9119

..

/etc/slurm/cluster=bar/slurm.conf

..

StateSaveLocation=/var/slurm-bar

SlurmdSpoolDir=/var/slurmd-bar

SlurmctldPidFile=/var/run/slurmctld-bar.pid

SlurmdPidFile=/var/run/slurmd-bar.pid

SlurmctldLogFile=/var/log/slurmctld-bar.log

SlurmdLogFile=/var/log/slurmd-bar.log

SlurmctldPort=9127

SlurmdPort=9128

AccountingStoragePort=9119

..

/etc/slurm/db=unionb/slurmdbd.conf

..

LogFile=/var/log/slurmdbd-unionb.log

PidFile=/var/run/slurmdbd-unionb.pid

DbdPort=9219

..

/etc/slurm/cluster=cluster1b/slurm.conf

..

StateSaveLocation=/var/slurm-cluster1b

SlurmdSpoolDir=/var/slurmd-cluster1b

SlurmctldPidFile=/var/run/slurmctld-cluster1b.pid

SlurmdPidFile=/var/run/slurmd-cluster1b.pid

SlurmctldLogFile=/var/log/slurmctld-cluster2b.log

SlurmdLogFile=/var/log/slurmd-cluster2b.log

SlurmctldPort=9217

SlurmdPort=9218

AccountingStoragePort=9219

..

/etc/slurm/cluster=cluster2b/slurm.conf

..

StateSaveLocation=/var/slurm-cluster2b

SlurmdSpoolDir=/var/slurmd-cluster2b

SlurmctldPidFile=/var/run/slurmctld-cluster2b.pid

SlurmdPidFile=/var/run/slurmd-cluster2b.pid

SlurmctldLogFile=/var/log/slurmctld-cluster2b.log

SlurmdLogFile=/var/log/slurmd-cluster2b.log

SlurmctldPort=9227

SlurmdPort=9228

AccountingStoragePort=9219

..

/etc/slurm/cluster=cluster3b/slurm.conf

..

StateSaveLocation=/var/slurm-cluster3b

SlurmdSpoolDir=/var/slurmd-cluster3b

SlurmctldPidFile=/var/run/slurmctld-cluster3b.pid

SlurmdPidFile=/var/run/slurmd-cluster3b.pid

SlurmctldLogFile=/var/log/slurmctld-cluster3b.log

SlurmdLogFile=/var/log/slurmd-cluster3b.log

SlurmctldPort=9237

SlurmdPort=9238

AccountingStoragePort=9219

The contents of the respective MSLURM-configuration ("mslurm.conf") allows for all MSLURM possible combinations as illustrated below:

MSLURM_SLURM_CONF="/etc/slurm/cluster=%cn/slurm.conf"

MSLURM_SLURMDBD_CONF="/etc/slurm/db=%un/slurmdbd.conf"

MSLURM_SLURM_CLUSTERNAMES_baz="foo bar"

MSLURM_SLURM_CLUSTERNAMES_unionb="cluster1b cluster2b cluster3b"

MSLURM_SLURM_CLUSTERNAMES="$MSLURM_SLURM_CLUSTERNAMES_baz $MSLURM_SLURM_CLUSTERNAMES_unionb"

MSLURM_SLURMDBD_UNIONNAMES="baz unionb"

Here the fixed assigned wildcards %cn and %un stand for a SLURM clusters or SLURM database names.

A SLURM cluster name must be identical to the contents of the variable cluster name of the respective SLURM configuration file.

8. Compute nodes

Also, you should verify that the participant SLURM nodes in one SLURM cluster (for example foo) receive the content of the corresponding SLURM configuration (here "/etc/slurm/cluster=foo/slurm.conf") in the standard path ("/etc/slurm/slurm.conf").

As there is also a "standard" SLURM configuration on the compute nodes, SLURM is also managed with the ordinary SLURM tools ("/etc/init.d/slurm, sinfo, scontrol...").

9. Examples

Daemons belonging to a single SLURM cluster or SLURM data base are managed at the MSLURM master via

mslurm <SLURM cluster> {start | status | stop | ..}

or

mslurmdbd <SLURM data base> {start | status | stop | ..}

Any SLURM commands to the SLURM cluster are executed with

mslurm <SLURM cluster> <SLURM-Command> {} <arguments>

All SLURM clusters of a SLURM data base can be addressed by the name of SLURM data base:

mslurm <SLURM data base> <SLURM-Command> {} <arguments>

All the objects can be combined with the option "-a" instead of specifying a SLURM cluster or SLURM data bases:

mslutm -a status

mslurmdbd -a status

mslurm -a sinfo

(The "-a" option is used by the MSLURM startup scripts when booting)

Also CVS lists of objects and actions can be executed:

mslurm foo,unionb scontrol show config

mslurm foo status,sinfo