LEIBNIZ RECHEN ZENTRUM

# MSLURM
# Multi cluster management

juan.pancorbo@lrz.de

Bayerische Akademie der Wissenschaften

# Introduction

- **LRZ is the computer centre for Munich's universities and is part of the Bavarian Academy of Sciences and Humanities.**
- **As a service provider for scientific high performance computing, LRZ operates compute systems for use by educational institutions in Munich, Bavaria, as well as on the national level.**
- **LRZ provides own computing resources as well as housing and managing computing resources from other institutions such as Max Planck Institute, Technische Universität München, or Ludwig Maximilians University.**

# Motivation

- **The tier 2 Linux cluster operated at LRZ is a heterogeneous system with different types of compute nodes, divided into 13 different partitions, each managed by SLURM.**
- **The various partitions are configured for the different needs and services requested, ranging from single node multiple core NUMAlink shared memory clusters, to a 16-way infiniband-connected cluster for parallel job execution, or an 8-way Gbit Ethernet cluster for serial job execution.**
- **The size of the clusters ranges are small in some cases (1,4,8 or 10 nodes) and up hundreds of nodes for others.**
- **Having one different host for each cluster where to run the slurm control daemon was not desired. A centralized solution was requested**

# Resources managed by slurm

| | Cluster name | Number of Nodes | CPU per Node | Connection |
|---|---|---|---|---|
| 1 | bsbslurm | 35 | 4 | Gbit Ethernet |
| 2 | capp | 15 | 8,48 | Gbit Ethernet |
| 3 | gpgpu | 4 | 16 | Gbit Ethernet |
| 4 | gvs | 4 | 16 | Gbit Ethernet |
| 5 | ice1 | 60 | 16 | Infiniband |
| 6 | inter | 8 | 16 | Infiniband |
| 7 | lmu_asc | 134 | 8,16 | Gbit Ethernet |
| 8 | mpp1 | 166 | 16 | Infiniband |
| 9 | myri | 10 | 32 | Myrinet |
| 10 | serial | 271 | 4,8 | Gbit Ethernet |
| 11 | tum_geodesy | 8 | 8 | Gbit Ethernet |
| 12 | uv2 | 1 | 1920 | NUMAlink |
| 13 | uv3 | 1 | 2240 | NUMAlink |

# What is MSLURM?

- **MSLURM is a script which allows you to run several slurm clusters (each one with his own configuration) from the same control node.**
- **MSLURM is essentially a wrapper which uses the environment variable $SLURM_CONF to specify a different slurm.conf file to each particular cluster.**
- **That means that different slurmctld daemons or slurmdbd daemons will be running on the same machine, and the system administrators can send slurm commands to any (or all) of the clusters in an easy way.**

# Use of MSLURM

- **The first parameter in the MSLURM call is the cluster name and the rest of the parameters are the standard slurm command with its own parameters.**
- **mslurm cluster_name sinfo**

```
lxs001:/etc/slurm # mslurm serial sinfo
PARTITION    AVAIL  TIMELIMIT NODES STATE  NODELIST
serial_std*  up    10-00:00:0   189 alloc  lx64a[301-489]
serial_large up    10-00:00:0    37 alloc  lx64a[133-169]
serial_long  up    20-00:00:0    20 alloc  lx64a[256-275]
serial_shm4  up    5-00:00:00     1 drain  lx64a290
serial_shm4  up    5-00:00:00    24 alloc  lx64a[276-289,291-300]
lxs001:/etc/slurm #
```

# MSLURM examples (I)

- **MSLURM can check the status of all the configured slurmctld**

```
lxs001:/etc/slurm # mslurm -a status
cluster inter: slurmctld (pid 13225) is running...
cluster ice1: slurmctld (pid 13262) is running...
cluster uv2: slurmctld (pid 13299) is running...
cluster uv3: slurmctld (pid 20265) is running...
cluster mpp1: slurmctld (pid 13373) is running...
cluster gvs: slurmctld (pid 13410) is running...
cluster myri: slurmctld (pid 13447) is running...
cluster gpgpu: slurmctld (pid 13484) is running...
cluster serial: slurmctld (pid 13519) is running...
cluster lmu_asc: slurmctld (pid 13558) is running...
cluster tum_geodesy: slurmctld (pid 13593) is running...
cluster bsbslurm: slurmctld (pid 13630) is running...
cluster capp: slurmctld (pid 7148) is running...
cluster rsrvd: slurmctld (pid 2086) is running...
cluster local: slurmctld (pid 7186) is running...
cluster local: slurmd (pid 7281) is running...
lxs001:/etc/slurm #
```

# MSLURM examples (II)

- **Also for all the configured slurmdbd**

```
lxs001:~ # mslurmdbd -a status
union cos: slurmdbd (pid 14932) is running...
union res: slurmdbd (pid 11746) is running...
union test: slurmdbd (pid 11518) is running...
union ruhl: slurmdbd (pid 11692) is running...
union lcgdb: slurmdbd (pid 23405) is running...
union pruebas: slurmdbd (pid 11726) is running...
lxs001:~ #
```

# MSLURM examples (III)

- **It can start and stop the daemons**

```
lxs001:~ # mslurm lcg stop;mslurm lcg start
cluster lcg: stopping slurmctld:
cluster lcg: slurmctld is stopped
cluster lcg: starting slurmctld:
lxs001:~ #
```

# MSLURM examples (IV)

- **The issued clusters can run with the same database or not.**

```
lxs001:~ # mslurm -a scontrol show conf | grep AccountingStoragePort
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6829
AccountingStoragePort   = 6959
AccountingStoragePort   = 6939
AccountingStoragePort   = 6949
AccountingStoragePort   = 6279
lxs001:~ #
```

# MSLURM examples (V)

- **It provides additional debug info if requested.**

```
lxs001:~ # mslurm -d lcg,rsrvd sinfo
mslurm: DEBUG: cluster lcg: objname='lcg'
mslurm: DEBUG: cluster lcg: export SLURM_CONF='/etc/slurm/cn=lcg/slurm.conf'
mslurm: DEBUG: cluster lcg: export SLURM_FILEEXT='-lcg'
mslurm: DEBUG: cluster lcg: export SLURM_MESGPRE='cluster lcg: '
mslurm: DEBUG: cluster lcg: export SLURM_CLUSTERNAME=lcg
mslurm: DEBUG: cluster lcg: command='sinfo'
PARTITION    AVAIL  TIMELIMIT NODES STATE  NODELIST
lcg_serial*  up    2-00:00:00     1 idle   lx64e14
mslurm: DEBUG: cluster rsrvd: objname='rsrvd'
mslurm: DEBUG: cluster rsrvd: export SLURM_CONF='/etc/slurm/cn=rsrvd/slurm.conf'
mslurm: DEBUG: cluster rsrvd: export SLURM_FILEEXT='-rsrvd'
mslurm: DEBUG: cluster rsrvd: export SLURM_MESGPRE='cluster rsrvd: '
mslurm: DEBUG: cluster rsrvd: export SLURM_CLUSTERNAME=rsrvd
mslurm: DEBUG: cluster rsrvd: command='sinfo'
PARTITION    AVAIL  TIMELIMIT NODES STATE  NODELIST
mpp1_rsrvd*  up      12:00:00     4 idle   lxa[175-178]
mpp1_double  up    1-00:00:00     4 idle   lxa[175-178]
lxs001:~ #
```

# MSLURM examples (VI)

- **It can show the parameters of the clusters so you can compare them by grepping the output**

```
lxs001:~ # mslurm -a scontrol show conf | grep DefMemPerCPU
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = 1500
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = 1000
DefMemPerCPU            = 1990
DefMemPerCPU            = 500
DefMemPerCPU            = 2000
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = 1000
DefMemPerCPU            = UNLIMITED
DefMemPerCPU            = 1990
lxs001:~ #
```

# MSLURM configuration

- **An additional mslurm.conf is needed, which is located at /etc/slurm. In this file you have to specify the cluster name and the database that it will use.**
- **The standard configuration files (slurm.conf gres.conf) are located on a subfolder named with the cluster name defined in mslurm.conf (/etc/slurm/cn=cluster_name)**

# Advantages

- **Centralized management. Is not needed a separate host for each cluster. All of them can run on the same host.**
- **Slurm commands can be sent to clusters on different databases with a single mslurm command.**
- **Parameters which are cluster-based but not partition-based (like DefMemPerCPU or SelectTypeParameters) can be added to a new different cluster with different values.**
- **Easy to add new clusters. Since slurm is already installed you only need to add the new slurm.conf file to the new folder, create the var-slurm folder (with the appropriate permissions), add the cluster to the mslurm.conf and add the cluster to the database.**

# Disdvantages

- **SPOF**
- **When changing the standard output of squeue with –o the spaces characters in the string must be escaped “%p\ %N\ %c”**
- **All the clusters have the same version of slurm**

```
lxs001:~ # mslurm -a sinfo -V
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
slurm 2.2.7
lxs001:~ #
```

# Questions

# Thank you!