SGI® Simple Linux® Utility for Resource Management Install Guide

007–5814–001

Record of Revision

Version	Description
001	November 2011 Original Printing.

Contents

About This Guide

This publication documents the SGI® implementation of the Simple Linux® Utility for Resource Management (SLURM) and how to install and configure it.

Related Publications and Additional Information

This section describes documentation you may find useful, as follows:

- *SGI Performance Suite 1.3 Start Here*

  Describes SGI Performance Suite 1.3 release including release features, software installation, and product overview. It includes a list of active SGI software and hardware manuals.

- *Message Passing Toolkit (MPT) User's Guide*

  Describes industry-standard message passing protocol optimized for SGI computers.

- *MPInside Reference Guide*

  Documents the SGI MPInside MPI profiling tool.

Online caveats, release note updates, and other useful information is available on the SGI SLURM pages on Supportfolio Online at the following location: https://support.sgi.com/

Note: You must be logged onto Supportfolio to access these pages.

Obtaining Publications

You can obtain SGI documentation in the following ways:

- See the SGI Technical Publications Library at: http://docs.sgi.com. Various formats are available. This library contains the most recent and most comprehensive set of online books, release notes, man pages, and other information.

- You can also view man pages by typing `man` *title* on a command line.

Conventions

The following conventions are used throughout this document:

Convention	Meaning
`command`	This fixed-space font denotes literal items such as commands, files, routines, path names, signals, messages, and programming language structures.
`manpage`(*x*)	Man page section identifiers appear in parentheses after man page names.
variable	Italic typeface denotes variable entries and words or concepts being defined.
`user input`	This bold, fixed-space font denotes literal items that the user enters in interactive sessions. (Output is shown in nonbold, fixed-space font.)
[ ]	Brackets enclose optional portions of a command or directive line.
...	Ellipses indicate that a preceding element can be repeated.

Reader Comments

If you have comments about the technical accuracy, content, or organization of this publication, contact SGI. Be sure to include the title and document number of the publication with your comments. (Online, the document number is located in the front matter of the publication. In printed publications, the document number is located at the bottom of each page.)

You can contact SGI in any of the following ways:

- Send e-mail to the following address:

  techpubs@sgi.com

- Contact your customer service representative and ask that an incident be filed in the SGI incident tracking system.
- Send mail to the following address:

SGI
Technical Publications
46600 Landing Parkway
Fremont, CA 94538

SGI values your comments and will respond to them promptly.

Introduction

This chapter describes the SGI® implementation of the Simple Linux® Utility for Resource Management (SLRUM) and covers these topics:

Simple Linux Utility for Resource Management

SGI® Simple Linux® Utility for Resource Management (SGI SLURM) is a fully supported SLURM release based on the community SLURM 2.2.7 offering that is available at the following site: https://computing.llnl.gov/linux/slurm/slurm.html.

SGI has added SGI MPI support as a standard SLURM MPI plugin called `sgimpi`. SGI MPI is the SGI MPI implementation, called Message Passing Toolkit (MPT). SGI MPI includes several additional software packages, such as, `memacct`, `sgi-array-services`, `sgi-mpt-shmem`, `xpmem`, and the MPInside profiling tool. For more information on SGI MPI, see Chapter 1. “Release Features” in the *SGI Performance Suite 1.3 Start Here*. For detailed information on MPT, see the *Message Passing Toolkit (MPT) User’s Guide*. These manuals are both available at the SGI Technical Publications Library at http://docs.sgi.com.

Note: The SGI MPI product is not included in the SLURM product offering. It must be purchased separately. The `sgimpi` plugin ships as part of the SGI SLURM product.

SGI SLURM is supported on SGI x86–64–based platforms running Red Hat Enterprise Server 6 (RHEL 6) or SUSE Linux Enterprise Server 11 (SLES 11) in clustered environments. Support for the SGI Altix UV platform is not currently available.

The official SLURM website (https://computing.llnl.gov/linux/slurm/slurm.html) describes SLURM as “an open-source resource manager designed for Linux clusters of all sizes.” It provides three key functions, as follows:

- “ It allocates exclusive and/or non-exclusive access to resources (computer nodes) to users for some duration of time so they can perform work.”

- “It provides a framework for starting, executing, and monitoring work (typically a parallel job) on a set of allocated nodes.”
- “It arbitrates contention for resources by managing a queue of pending work."

For SLURM installation and configuration information, see Chapter 2, "Simple Linux Utility for Resource Management Installation and Configuration" on page 3.

SLURM RPM Information and Release Notes

For a complete list of RPMs included in the SGI SLURM release, see the file called RPMS.txt that is available in the `/docs` directory on the CD media for the SGI SLURM product.

For the latest information about software and documentation in this release, see the release notes that are in a file with the product name and -readme.txt suffix that is available in `/docs` directory on the CD media for the SGI SLURM product.

Simple Linux Utility for Resource Management Installation and Configuration

This chapter describes how to install SGI® Simple Linux Utility for Resource Management (SGI® SLURM). It covers the following topics:

Installing the SLURM RPMs

Depending on the type of clustered environment, the SLURM RPMs should be installed, as follows:

- SGI Altix ICE systems
  - service node: install all SLURM RPMs
  - compute nodes: install all SLURM RPMs except `slurm-slurmdb-direct`
- Flat clusters, such as, SGI® Rackable™ servers
  - login node: install all SLURM RPMs
  - compute nodes: install all SLURM RPMs except `slurm-slurmdb-direct`

Note: Do **NOT** start any `munge`SLURM daemons yet. Instructions for configuring the `munge` and SLURM daemons are provided in the following two sections.

Configuring munge Authentication Plugin

SLURM uses the munge AuthType plugin for default authentication. SGI strongly recommends use of the munge plugin, however, other authentication methods are available. See the slurm.conf(5) man page at section "AuthType" for more options.

If you do not plan to use munge authentication, you do not need to configure munge and may skip the rest of this section.

On a service, login or head nodes, perform the following:

1.

   ```
   % chkconfig munge on
   ```

   Change the OOM value in /etc/sysconfig/munge to -17 (recommended)

2. Generate the munge key, as follows:

   ```
   % echo -n "foo" | sha1sum | cut -d' ' -f1 >/etc/munge/munge.key
   % chown -R daemon /etc/munge/munge.key
   % chmod -R 600 /etc/munge/munge.key
   ```

3.

   ```
   % /etc/init.d/munge restart
   ```

4.

   ```
   % scp -p /etc/munge/munge.key <each_other_node>:/etc/munge
   ```

5.

   ```
   % scp -p /etc/sysconfig/munge <each_other_node>:/etc/sysconfig
   ```

6. On an SGI Altix ICE system, you can use the cpush command for steps 3 and 4, above, as follows:

   ```
   % cpush /etc/munge/munge.key /etc/munge
   % cpush /etc/sysconfig/munge /etc/sysconfig
   ```

   For more information on the cpush command, see *SGI Management Center for Altix ICE*.

On each other node type, perform the following:

1.
```
% chkconfig munge on
```
2.
```
% /etc/init.d/munge restart
```

Configuring SLURM

Configuration information for SLURM is available in several locations, as follows:

- Examples are in the `/etc/slurm` directory.
- A configuration helper is available at this site:
  http://www.llnl.gov/linux/slurm/configurator.html
- Overall configuration parameters are described in the `slurm.conf`(5) man page.

Changes to the SLURM configuration should be made in `/etc/slurm/slurm.conf` file. SGI recommends some default configuration parameters, as follows:

```
JobAcctGatherType=jobacct_gather/linux
TaskPlugin=task/affinity
MpiDefault=sgimpi
AccountingStorageType=accounting_storage/slurmdbd
ClusterName=george
```

Various SLURM configurations require the user to create directories. The notes below reflect some of the directories that some SLURM configurations often require. You should tailor the information provided below to your specific site policies and practices, as follows:

- All SLURM `Nodename` specifications are based on the output from the `hostname -s` command.
- Do not use a hostname alias for `ControlMachine`. Use `service0` for example.
- For `NodeName` configuration, you can use a hostname alias if you also use the `NodeHostname` option as shown in the following example:

  ```
  NodeName=db[1-4] NodeHostname=dewberry[1-4] ...
  ```

Partition Nodes reference NodeName

```
Partition=... Nodes=db[1-2] ...
```

- NodeName RealMemory Units are in MBytes, so divide the result by 1024, as follows:

```
% head -1 /proc/meminfo  | awk '{printf "%.0f\n", $2/1024;}'
```

On SGI Altix ICE systems, take the lowest value or write an entry for each node, as follows:

```
% cexec --all --pipe  head -1 /proc/meminfo  | \
        awk '{printf "%.0f\n", $4/1024;}
```

- NodeName CPU counts are easily obtained via lk_hostid, as follows:

```
SLURM specification: Sockets=2 CoresPerSocket=4 ThreadsPerCore=1 Procs=8
```

which corresponds to the following commands:

```
% lk_hostid -1CC     : socket=2  core=8   processor=8
```

On SGI Altix ICE systems, perform the following:

```
% cexec --all --pipe lk_hostid -1CC
```

- Create /var/log/slurm on the service node. /var/log/slurm corresponds to the directory entries in slurm.conf and slurmdbd.conf. For example:
  - in slurm.conf: SlurmctldLogFile=/var/log/slurm/slurm.log
  - in slurmdbd.conf: ArchiveDir=/var/log/slurm/accounting_archive and LogFile=/var/log/slurm/slurmdbd.log

  Perform these commands:

```
% mkdir -p /var/log/slurm
% chmod 775 /var/log/slurm
% chgrp slurm /var/log/slurm
```

- Make sure that StateSaveLocation points to a shared directory accessible to the service, login or head node and all compute nodes. The following examples assume that StateSaveLocation=/data/slurm.

On service, login or head nodes, create the shared directory and setup NFS using the following commands and/or mount examples:

```
% mkdir -p /data/slurm
% chmod 775 /data/slurm
% chgrp slurm /data/slurm
```

Add mount points. Mount points are set differently for SGI Altix ICE and flat clusters, as follows:

– For SGI Altix ICE systems:

```
service0-ib1:/data/slurm   /data/slurm  nfs rw,intr,hard 0
```

– For flat clusters:

```
<serviceNode>:/data/slurm  /data/slurm  nfs rw,intr,hard 0 0
```

- Configure the `TmpFS` directory

  SLURM sets `TmpFS=/tmp` by default in the `slurm.conf` file. SLURM uses it to store various temporary files. Users should consider that this directory could be the current directory when a SLURM job is launched. Hence, should a job core dump for example, there should be sufficient space to capture the dump. Since SLURM monitors `TmpFS` directory as a resource, if it becomes full (because of a core dump, for example), SLURM will make the node unavailable until the problem is fixed. As a result, the node will not be available for further job processing.

  For this reason, it is a very important configuration item.

  On an SGI Altix ICE cluster, `/tmp` is usually very small (that is, 150 Mbytes) by default. It is likely not sufficient. SGI strongly recommends that TmpFS **NOT** be set to `/tmp`. Instead, use another mount point, such as, an NFS mount point. For example, `/data/slurm/tmp` from the service node. See the prior bullet item.

  The `TmpFS` directory permissions should be `1777`.

Configuring `mysql` and `slurmdbd`

This section provides additional notes on configuring `mysql` and `slurmdbd`. Three issues to note, are as follows:

- slurmdbd configuration is only required when the following value is set in the slurm.conf configuration file:

  AccountingStorageType=accounting_storage/slurmdbd

- The slurmdbd service may be installed on certain node types within a cluster or on a system outside of the cluster. On SGI Altix ICE systems, SGI recommends that the slurmdbd service be installed/configured to run on the service node. On flat clusters, SGI recommends that the slurmdbd service be installed/configured to run on a login node if you have one or the head node if you want to run within the cluster or on an external system.
- The slurmdbd service does **NOT** need to be configured and active on compute nodes.

Install MySQL

Install MySQL if it is not already installed with the following yum or zypper command(s):

```
% yum install mysql-server            # on RHEL 6 systems
% zypper install mysql                # on SLES 11 systems
```

Note: SGI recommends installing MySQL on the same node (or system) where slurmdbd will be installed and configured.

Start MySQL

It is required that you start MySQL, as follows:

```
% /etc/init.d/mysqld restart          # on RHEL 6 systems
% /etc/init.d/mysql restart           # on SLES 11 systems
```

Configure MySQL

For information on configuring MySQL for SLURM, see https://computing.llnl.gov/linux/slurm/accounting.html.

Invoke mysql, as root user, using one of the following methods:

- On SGI Altix ICE systems managed with SGI Management Center (SMC), MySQL should already be installed, so invoke it using the MySQL root password stored in /etc/odapw, as follows:

```
% mysql --password=‘cat /etc/odapw‘
```

- For flat clusters or SGI Altix ICE systems which do not have a MySQL root password set, perform the following:

```
% mysql
```

- On systems where the system administrator has set a root MySQL password, use that password, as follows:

```
% mysql --password=<MySQL root password>
```

To create user slurm and assign a password abc123 that is the password specified in the slurmdbd.conf configuration file, perform the following:

```
mysql> create user slurm IDENTIFIED BY ’abc123’;
```

To grant all privileges to user slurm, perform the following:

```
 mysql> grant all on slurm_acct_db.* TO ’slurm’@’localhost’
identified by ’abc123’ with grant option;
```

Configure /etc/slurm/slurmdbd.conf File

For detailed recommendations on the slurmdbd.conf configuration file, see "Configuring SLURM" on page 5.

Note: The ClusterName entry is **extremely important**. Be careful with your choice and do not change.

Create the slurm accounting_archive data base, as follows:

```
% mkdir -p  /var/log/slurm/accounting_archive
```

To start the slurmdbd service, perform the following:

```
% chkconfig -add slurmdbd
% chkconfig slurmdbd on
```

```
% /etc/init.d/slurmdbd start
```

At a minimum, add your cluster and an account to the slurmdbd service with the following commands:

```
% sacctmgr add cluster <ClusterName value in slurm.conf>
% sacctmgr add account none,test Cluster=<ClusterName value in slurm.conf> \
Description="none" Organization="none"
```

For more information slurmdbd configuration options, see "Configuring SLURM" on page 5.

Starting SLURM

To start SLURM, perform the following commands on the cluster nodes specified below. The last series of commands in this section is for SGI Altix ICE systems.

If you plan to use **slurmdbd**, restart the slurmdbd service on the service, login or head node, as follows:

```
% /etc/init.d/slurmdbd restart
```

Start the slurm daemon on all nodes, including the service node, as follows:

```
% /etc/init.d/slurm restart
```

On SGI Altix ICE systems, run the following commands from the service node:

```
% /etc/init.d/slurmdbd restart (if you plan to use slurmdbd)
% /etc/init.d/slurm restart
% cexec --all --pipe /etc/init.d/slurm restart
```

Basic Test Output Example

This section provides some sample output of a non-MPT based job example.

```
# salloc -N2 -n4   # 2= number of nodes , 4=total number of CPUs on all nodes
salloc: Granted job allocation 11

# srun -l /bin/hostname
0: dewberry3
1: dewberry3
```

```
2: dewberry4
3: dewberry4

# srun -l date
0: Tue Aug  2 18:44:59 CDT 2011
1: Tue Aug  2 18:44:59 CDT 2011
2: Tue Aug  2 18:44:59 CDT 2011
3: Tue Aug  2 18:44:59 CDT 2011
# exit
salloc: Relinquishing job allocation 11

# sacct
       JobID    JobName  Partition    Account  AllocCPUS      State ExitCode
------------ ---------- ---------- ---------- ---------- ---------- --------
11                 bash slurm-db-+       root          4  COMPLETED 0:0
11.0           hostname                  root          4  COMPLETED 0:0
11.1               date                  root          4  COMPLETED 0:0
```

Useful SLURM Commands

This section provides some commands you may find useful when using SLURM.

The following commands are usually run only from the service, login or head node as root:

- Node status:

```
scontrol show node              # Watch for State
sinfo -el                       # Watch for State
```

- Node status change:

```
scontrol updade Nodename=r1i3n[0-3] State=IDLE
```

- Partition status:

```
scontrol show partition    # Watch for State
```

- Partition state change:

  ```
  scontrol update part=partitionName state=IDLE
  ```

- Job status:

  ```
  sjstat
  ```

- Accounting

  ```
  sacct
  ```

Default Usernames and Password

When the SGI SLURM RPMs are installed, a user called "`slurm`" is created. The login shell value is set to `/bin/false`, no password is set for user `slurm` and the HOME directory is set to `/opt/sgi/slurm`.

Index