Oracle Grid Engine

Project Report

Spring 2011
IIIT, Hyderabad

Mayank Juneja (200702022)

Contents

1 Introduction

Oracle Grid Engine is a workload manager (job scheduler) which schedules jobs to run in a cluster environment. The main purpose of a job scheduler is to utilize system resources (CPU, memory, etc) in the most efficient way possible.

This kind of a job scheduler is useful in computationally intensive environments where tasks like running calculations, are usually done in a more "on demand" fashion. When a user needs something, he tells the server, and the server does it. For the most part it doesn't matter on which particular machine the calculations are run. All that matters is that the user can get the results. This kind of work is often called batch, offline, or interactive work. Sometimes batch work is called a job. Typical jobs include rendering images or movies, running simulations, processing input data, modeling chemical or mechanical interactions, and data mining.

Many organizations have hundreds, thousands, or even tens of thousands of machines devoted to running jobs.

The role of a workload manager is to take a list of jobs to be executed and distribute them across the available machines. The workload manager makes life easier for the users because they don't have to track all their jobs themselves, and it makes life easier for the administrators because they don't have to manage users' use of the machines directly. It's also better for the organization in general because a workload manager will usually do a much better job of keeping the machines busy than users would on their own, resulting in much higher utilization of the machines. Higher utilization effectively means more compute power from the same set of machines, which means faster results and higher capacity without having to purchase additional resources.

2 Basics

2.1 Types of Hosts

Master Host The master host is the central component of an Oracle Grid Engine compute cluster. It is responsible for accepting incoming jobs from users, assigning jobs to resources, monitoring the overall cluster status , and processing administrative commands. It runs a multi-threaded daemon that runs on a single host in the compute cluster.

Shadow Master Host Apart from the master host, there can be shadow master(s) that can take over when the master host fails. The presence of shadow master(s) is optional and helps in reducing the downtime.

Execution Host An execution host is a system in the cluster where the jobs get executed (actual computations are carried out on the execution hosts locally).

Submit Host A sumbit host is a system from where users can submit jobs.

Admin Host An admin host is a system where Grid Engine administrative commands can be run.

2.2 Queues

2.3 Supported Platforms

Master Host

1. Solaris
2. Linux kernel 2.4-2.6 on x86/x64 (glibc at least 2.3.2)

Execution Host

1. Solaris
2. Linux kernel 2.4-2.6 on x86/x64 (glibc at least 2.3.2)
3. Microsoft Windows
4. AIX
5. HP-UX

3 Installation

3.1 Installation Layout

We will describe the installation procedure on CentOS. The procedure will remain almost similar on other Unix based Operating Systems. For describing the installation procedure, we will assume the following setup :

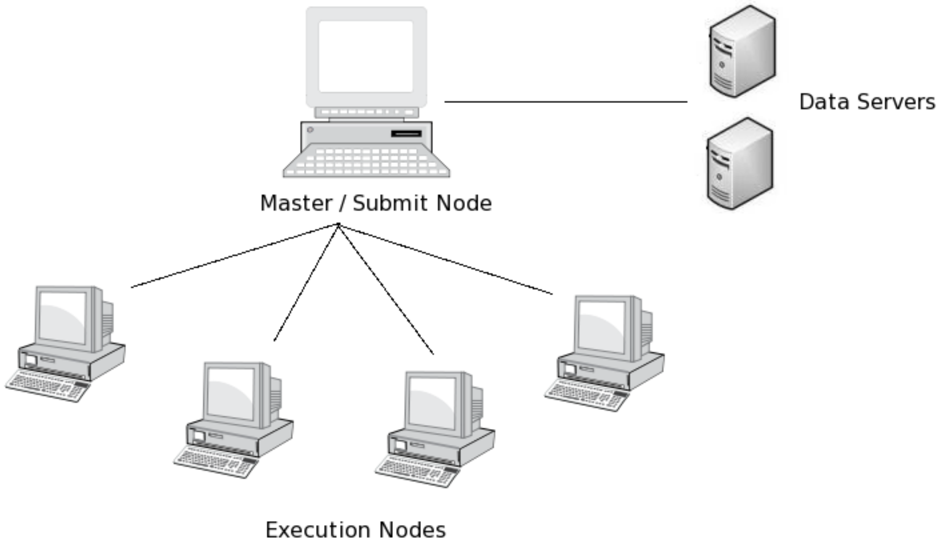

Figure 1: Cluster Setup

- A central node serving as Master host. It will also serve as submit host. Let the IP of the master host be 10.1.1.10.
- Multiple exeecution hosts. Let the IPs of the execution hosts be 10.1.1.11-10.1.1.20
- Data servers. The data servers are mounted on the master host and the execution hosts. The data servers will store the user's data. The need of independent data servers occur because of the huge data size of the users in the cluster environment. Let the IPs of the data server be 10.1.1.21.

3.1.1 SGE_ROOT Directory

The directory in which the Oracle Grid Engine software is unpacked is known as the root or SGE_ROOT (because the directory path is stored in the $SGE_ROOT environment variable). The root directory contains all of the files required to run the Oracle Grid Engine software on one or more binary architectures. The root directory is self-contained, meaning that the installation process does not create or modify files anywhere other than in the root directory. We will create the SGE_ROOT on the master node and mount it across the execution nodes.

In order to support multiple binary architectures from the same root directory, the bin and lib directories each have architecture-specific subdirectories. For example, on a Unix 64 bit host, the path to the qsub command will be $SGE_ROOT/bin/lx24-amd64/qsub.

The installation process will create a new cell (or SGE_CELL) directory in the root directory. The cell directory contains all of the files associated with the installation. Because the cell directory is completely self-contained, multiple cells can coexist with in the same root directory, with each cell representing a different Oracle Grid Engine installation. As long as the cells do not share port numbers or spool directory paths, all cells under a given root may be simultaneously active, i.e. they may each have a running set of daemons associated with them. In our case, we have only one cell.

Each cell directory contains a common directory and a spool directory. The common directory contains configuration files, utility scripts, and some cluster data files. The spool directory contains some or all of the daemon log (messages) files and may contain the qmaster's data store. Most importantly, the common directory inside the cell directory contains the files used by clients to find the qmaster. When a client attempts to contact the qmaster, it reads the $SGE_ROOT/ $SGE_CELL/common/act_qmaster file to find out on which host the qmaster currently resides. (Because of fail-over, the qmaster may switch hosts periodically.) For this reason, it is important that the cell directory be shared with all potential client hosts, including the execution hosts.

3.1.2 QMaster Data Spooling

The majority of the configuration information for an Oracle Grid Engine cluster, including the complete data for all active jobs, is stored in the qmaster's data spool. The data spool ensures that the state of the cluster is persistent between qmaster restarts. The Oracle Grid Engine software offers three options for the type of data spool as well as the choice of where the data spool should be stored.

The qmaster's data spool can be based on flat files (known as classic spooling because it was once the only option), a local Berkeley Database data store, or a remote Berkeley Database server. The computing environment, the size of the cluster, and the expected cluster throughput are key factors used in determining which spooling option is best. Before making any decision based on performance expectations, one should always do performance testing in the local compute environment to determine the best solution.

3.1.3 Execution Daemon Data Spooling

Each execution daemon has its own spool directory. The default location for the spool directory offered by the installation process is $SGE_ROOT/$SGE_CELL/spool/<hostname>.

The execution daemon's spool directory contains information about the jobs currently being executed by that execution daemon, so that in the event of a failure, the execution daemon can try to reconstruct its previous state when it restarts. Among the information stored in the execution daemon's spool directory is a local copy of every job script being executed by that execution daemon. When a user submits a binary job, only the path to the binary is passed on to the execution daemon. When a user submits a script job, however, the script file is copied and sent along with the job. When the execution daemon executes the job, it runs a local copy of the script saved into its spool directory, rather than the original script file submitted by the user.

3.1.4 Managing User Data

The Oracle Grid Engine software does not manage user data by default. Whatever data is needed by users' jobs must already be accessible from the execution hosts, normally via some kind of shared or distributed file system, such as NFS or the Lustre file system. It is very important to plan effectively for the handling of user data. Poorly planned data management can bring even the best cluster to its knees.

A shared or distributed file system is the most common approach to data availability. The advantage of a shared or distributed file system is that every job sees the same set of data, and there are no data synchronization issues.

Shared and distributed file systems introduce their own set of problems. The quality of service of an NFS file system tends to decay rapidly as the data throughput increases. Lustre, on the other hand, has excellent scalability characteristics, but is a heavyweight solution that requires a significant amount of administrative attention. There are other alternatives, such as QFS or pNFS, and each comes with its own set of advantages and limitations. If a shared or distributed file system is chosen as the approach to data availability, be sure to examine all the options and pick one that is appropriate for the computing environment and the needs of the jobs.

3.1.5 Naming Services

The Oracle Grid Engine software is relatively agnostic to the naming service used in the compute environment. The most important thing is that the master host be able to resolve the host names of all the execution hosts and client hosts, and that all client hosts and execution hosts be able to resolve the name of the master host. Two host name resolution configuration options are provided to simplify matters even further.

During the installation process, the installing user will be asked if all of the hosts in the cluster share the same domain name. If the user answers 'yes,' the Oracle Grid Engine software will ignore domain names completely by truncating every host name at the first dot ('.'). If the user answers 'no,' the Oracle Grid Engine software will always use fully qualified host names.

3.1.6 User Accounts

When an execution daemon executes a job, it executes the job as the user who submitted the job. To enable that functionality, the name of the user who submitted the job is included in the job's definition. It is important to note that it is the name of the user who submitted the job, not the id, that is passed to the execution daemon. In order to execute the job as that user, that user's name must exist as a user account on the execution host. The execution daemon relies on the user management facilities of the underlying OS to resolve the user name to a local account. How the user name is resolved, be it through `/etc/passwd`, NIS or LDAP or some other mechanism, plays no role, as long as the user name can be resolved to a local account.

3.1.7 Admin Accounts

To enable the use of a privileged port and the ability to execute jobs as other users, the Oracle Grid Engine daemons must be started as the root user. It is actually possible to install and run a cluster as a non-root user, but then only that user will be able to run jobs in the cluster. To limit the security impact of the daemons running as root, they instead switch to a non-root user immediately after start- up. That non-root user can be specified during installation, and is by convention chosen to be `sgeadmin`. If the cluster is configured to use a non-root user in this way, it is important that the user chosen have an account on the master machine and every execution host.

3.1.8 Service Ports

An Oracle Grid Engine cluster requires two port numbers in order to function: one for the qmaster, and one for the execution daemons. Those port numbers can be assigned in one of two places, chosen during the cluster installation process. One option is to read the port numbers from the `sge_qmaster` and `sge_execd` entries in the `/etc/services` file. The other is to set the port numbers in the shells of client users.

3.2 Prerequisites

3.2.1 Setup a NIS server

NIS server is used to have a central authentication server for the end users. We can use other services like LDAP also for the purpose. We will setup the NIS serer on the master node itself. `ypserv` package is required for setting up NIS server.

- **Edit `/etc/sysconfig/network` File**

  Add the NIS domain you wish to use in the `/etc/sysconfig/network` file. Let us call the domain as GRID-ENGINE.

  Listing 1: Editing `/etc/sysconfig/network`

  ```
  #/etc/sysconfig/network
  NISDOMAIN="GRID-ENGINE"
  ```

- **Edit `/etc/yp.conf` File**

  NIS servers also have to be NIS clients themselves, so you'll have to edit the NIS client configuration file /etc/yp.conf to list the domain's NIS server as being the server itself or localhost.

  Listing 2: Editing `/etc/yp.conf`

  ```
  # /etc/yp.conf - ypbind configuration file
  ypserver 127.0.0.1
  ```

- **Start The Key NIS Server Related Daemons**

  Start the necessary NIS daemons in the /etc/init.d directory and use the `chkconfig` command to ensure they start after the next reboot.

  Listing 3: Starting NIS Server Daemons

  ```
  [root@master ]# service portmap start
  Starting portmapper: [  OK  ]
  [root@master ]# service yppasswdd start
  Starting YP passwd service: [  OK  ]
  [root@master ]# service ypserv start
  Setting NIS domain name GRID-ENGINE:  [  OK  ]
  Starting YP server services: [  OK  ]

  [root@master ]# chkconfig portmap on
  [root@master ]# chkconfig yppasswdd on
  [root@master ]# chkconfig ypserv on
  ```

- **Initialize Your NIS Domain**

  Now that you have decided on the name of the NIS domain, you'll have to use the `ypinit` command to create the associated authentication files for the domain. You will be prompted for the name of the NIS server, which in this case is `master`. With this procedure, all nonprivileged accounts are automatically accessible via NIS.

  Listing 4: Initializing NIS Domain

  ```
  [root@master ]# /usr/lib/yp/ypinit -m
  At this point, we have to construct a list of the hosts which will run NIS
  servers.  master is in the list of NIS server hosts.  Please continue to add
  the names for the other hosts, one per line.  When you are done with the
  list, type a <control D>.
    next host to add:  master
    next host to add:
  The current list of NIS servers looks like this:

  master

  Is this correct?  [y/n: y]  y
  ```

```
We need a few minutes to build the databases...
Building /var/yp/GRID-ENGINE/ypservers...
Running /var/yp/Makefile...
gmake[1]: Entering directory `/var/yp/GRID-ENGINE'
Updating passwd.byname...
Updating passwd.byuid...
Updating group.byname...
Updating group.bygid...
Updating hosts.byname...
Updating hosts.byaddr...
Updating rpc.byname...
Updating rpc.bynumber...
Updating services.byname...
Updating services.byservicename...
Updating netid.byname...
Updating protocols.bynumber...
Updating protocols.byname...
Updating mail.aliases...
gmake[1]: Leaving directory `/var/yp/GRID-ENGINE'

master has been set up as a NIS master server.

Now you can run ypinit -s master on all slave server.
[root@master ]#
```

- **Start The `ypbind` and `ypxfrd` Daemons**

  Start the necessary NIS daemons in the `/etc/init.d` directory and use the `chkconfig` command to ensure they start after the next reboot.

  Listing 5: Starting `ypbind` and `ypxfrd` Daemons

  ```
  [root@master ]# service ypbind start
  Binding to the NIS domain: [  OK  ]
  Listening for an NIS domain server.
  [root@master ]# service ypxfrd start
  Starting YP map server: [  OK  ]
  [root@master ]# chkconfig ypbind on
  [root@master ]# chkconfig ypxfrd on
  ```

3.2.2 Configure NIS clients

We have to configure NIS clients on the execution hosts. The following procedure needs to be carried out on all the execution hosts.

- **Run `authconfig`**

  The `authconfig` or the `authconfig-tui` program automatically configures the NIS files after prompting for the IP address and domain of the NIS server.

  Listing 6: Running `authconfig`

  ```
  [root@node1 ]# authconfig-tui
  ```

  Once finished, it creates an {texttt/etc/yp.conf file that defines, amongst other things, the IP address of the NIS server for a particular domain. It also edits the `/etc/sysconfig/network` file to define the NIS domain to which the NIS client belongs.

  Listing 7: Client's `/etc/yp,conf`

  ```
  # /etc/yp.conf - ypbind configuration file
  domain GRID-ENGINE server 10.1.1.10
  ```

Listing 8: Client's /etc/sysconfig/network

```
#/etc/sysconfig/network
NISDOMAIN=GRID-ENGINE
```

In addition, the authconfig program updates the /etc/nsswitch.conf file that lists the order in which certain data sources should be searched for name lookups, such as those in DNS, LDAP, and NIS. Here you can see where NIS entries were added for the important login files.

Listing 9: Client's /etc/nsswitch.conf

```
#/etc/nsswitch.conf
passwd:     files nis
shadow:     files nis
group:      files nis
```

- **Start The NIS Client Related Daemons**

  Start the ypbind NIS client, and portmap daemons in the /etc/init.d directory and use the chkconfig command to ensure they start after the next reboot.

Listing 10: Starting The NIS Client Related Daemons

```
[root@node1 ]# service portmap start
Starting portmapper: [  OK  ]
[root@node1 ]# service ypbind start
Binding to the NIS domain:
Listening for an NIS domain server.
[root@node1 ]#

[root@node1 ]# chkconfig ypbind on
[root@node1 ]# chkconfig portmap on
```

3.2.3 Configure the NFS Server

We will setup the NFS serer on the data servers. Let the partitions to be shared be /data1, /data2, ...

- **Edit the /etc/exports file**

  Edit the file to allow NFS mounts of the /data1, /data2, ... directories with read/write access.

Listing 11: Editing /etx/exports

```
/data1                  *(rw,sync,no_root_squash)
/data2                  *(rw,sync,no_root_squash)
```

- **Run exportfs**

  Let NFS read the {texttt/etc/exports file for the new entry, and make /data1, /data2, ... available to the network with the exportfs command.

Listing 12: Running exportfs

```
[root@data ]# exportfs -a
[root@data ]#
```

- **Make sure the required nfs, nfslock, and portmap daemons are both running and configured to start after the next reboot.**

Listing 13: Running NFS Daemons

```
[root@data ]# chkconfig nfslock on
[root@data ]# chkconfig nfs on
[root@data ]# chkconfig portmap on
[root@data ]# service portmap start
Starting portmapper: [  OK  ]
[root@data ]# service nfslock start
Starting NFS statd: [  OK  ]
[root@data ]# service nfs start
Starting NFS services:  [  OK  ]
Starting NFS quotas: [  OK  ]
Starting NFS daemon: [  OK  ]
Starting NFS mountd: [  OK  ]
[root@data ]#
```

3.2.4 Configure the NFS Clients

We will setup the NFS clients on the master host and the execution hosts. The following procedure needs to be carried out on all the execution hosts and the master host.

- **Make sure the required `netnfs`, `nfslock`, and `portmap` daemons are both running and configured to start after the next reboot.**

Listing 14: Running NFS Daemons

```
[root@node1 ]# chkconfig nfslock on
[root@node1 ]# chkconfig netfs on
[root@node1 ]# chkconfig portmap on
[root@node1 ]# service portmap start
Starting portmapper: [  OK  ]
[root@node1 ]# service netfs start
Mounting other filesystems:  [  OK  ]
[root@node1 ]# service nfslock start
Starting NFS statd: [  OK  ]
[root@node1 ]#
```

- **Create the directories**

Listing 15: Creating directories

```
[root@node1 ]# mkdir /data1
[root@node1 ]# mkdir /data2
```

- **Edit the `/etc/fstab`**

  Edit the `/etc/fstab` to define the mount points.

Listing 16: Editing `/etc/fstab`

```
10.1.1.21:/data1                /data1          nfs     rw,hard,intr    0       0
10.1.1.21:/data2                /data2          nfs     rw,hard,intr    0       0
```

3.2.5 Enable HostBased Authentication

Oracle Grid Engine assumnes password-less login from the submit/master host to the executioan hosts. This is required for submitting jobs. We will use HostBased Authentication to fulfill the requirement.

- **Configure the ssh server**

  The following settings need to be done on all the execution hosts (All of them are the ssh servers).

Listing 17: Add the following line in `/etc/ssh/sshd_config`

```
#/etc/ssh/sshd_config
HostbasedAuthentication yes
```

Listing 18: Add the client hostname in `/etc/hosts.equiv`

```
#/etc/hosts.equiv
master
```

- **Configure the ssh client**

  The following settings need to be done on the master/submit host.

Listing 19: Add the following line in `/etc/ssh/ssh_config`

```
#/etc/ssh/ssh_config
HostbasedAuthentication yes
```

Listing 20: Make the `ssh` executable setuid root

```
[root@master ]# chmod 4755 /usr/bin/ssh
[root@master ]#
```

Listing 21: Add the client hostname in `/etc/hosts.equiv`

```
#/etc/hosts.equiv
master
```

Add the server's ssh host key into `/etc/ssh/ssh_known_hosts`. Just paste the contents of server's `/etc/ssh/ssh_host_dsa_key.pub`.

3.3 Installing Oracle Grid Engine

3.3.1 Master Host

Assume the `SGE_ROOT` directory is `/grid-engine/ge6.2u5`. The following steps are to be carried out on the master host.

- **Go to `SGE_ROOT` directory**

Listing 22: Go to `SGE_ROOT` directory

```
[root@master ~]# cd /grid-engine/ge6.2u5/
[root@master ge6.2u5]#
```

- **Run `install_qmaster`**

Listing 23: Run `install_qmaster`

```
[root@master ~]# ./install_qmaster
```

  Accept the agreement.

- **Choose Grid Engine admin user account**

Listing 24: Choose Grid Engine admin user account

```
Choosing Grid Engine admin user account
---------------------------------------

You may install Grid Engine that all files are created with the user id of an
unprivileged user.

This will make it possible to install and run Grid Engine in directories
where user >root< has no permissions to create and write files and directories.

  - Grid Engine still has to be started by user >root<

- this directory should be owned by the Grid Engine administrator

Do you want to install Grid Engine
under an user id other than >root< (y/n) [y] >>
```

Press n.

- **Checking `$SGE_ROOT` directory**

Listing 25: Checking `$SGE_ROOT` directory

```
Checking $SGE_ROOT directory
----------------------------

The Grid Engine root directory is:

  $SGE_ROOT = /grid-engine/ge6.2u5

If this directory is not correct (e.g. it may contain an automounter
prefix) enter the correct path to this directory or hit <RETURN>
to use default [/grid-engine/ge6.2u5] >>
```

Hit <RETURN>.

- **Configure Grid Engine TCP/IP communication service**

Listing 26: Configure Grid Engine TCP/IP communication service

```
Grid Engine TCP/IP communication service
----------------------------------------

The port for sge_qmaster is currently set by the shell environment.

  SGE_QMASTER_PORT = 6444

Now you have the possibility to set/change the communication ports by using the
>shell environment< or you may configure it via a network service, configured
in local >/etc/service<, >NIS< or >NIS+<, adding an entry in the form

    sge_qmaster <port_number>/tcp

to your services database and make sure to use an unused port number.

How do you want to configure the Grid Engine communication ports?

Using the >shell environment<:                           [1]

Using a network service like >/etc/service<, >NIS/NIS+<: [2]

(default: 1) >>
```

Press 1.

- **Configure Grid Engine Cells**

Listing 27: Configure Grid Engine cells

```
Grid Engine cells
```

```
-----------------

Grid Engine supports multiple cells.

If you are not planning to run multiple Grid Engine clusters or if you don't
know yet what is a Grid Engine cell it is safe to keep the default cell name

   default

If you want to install multiple cells you can enter a cell name now.

The environment variable

  $SGE_CELL=<your_cell_name>

will be set for all further Grid Engine commands.

Enter cell name [default] >>
```

Press Enter.

- **Configure Cluster name**

Listing 28: Configure Cluster name

```
Unique cluster name
-------------------

The cluster name uniquely identifies a specific Sun Grid Engine cluster.
The cluster name must be unique throughout your organization. The name
is not related to the SGE cell.

The cluster name must start with a letter ([A-Za-z]), followed by letters,
digits ([0-9]), dashes (-) or underscores (_).

Enter new cluster name or hit <RETURN>
to use default [atom] >>
```

Enter cluster name and hit <RETURN>.

- **Configure Grid Engine qmaster spool directory**

Listing 29: Configure Grid Engine qmaster spool directory

```
Grid Engine qmaster spool directory
-----------------------------------

The qmaster spool directory is the place where the qmaster daemon stores
the configuration and the state of the queuing system.

User >root< on this host must have read/write access to the qmaster
spool directory.

If you will install shadow master hosts or if you want to be able to start
the qmaster daemon on other hosts (see the corresponding section in the
Grid Engine Installation and Administration Manual for details) the account
on the shadow master hosts also needs read/write access to this directory.

Enter a qmaster spool directory [/grid-engine/ge6.2u5/default/spool/qmaster] >>
```

Hit <RETURN>.

- **Configure Windows Execution host**

Listing 30: Configure Windows Execution Host

```
Windows Execution Host Support
------------------------------

Are you going to install Windows Execution Hosts? (y/n) [n] >>
```

If there are any Windows Execution host in the cluster, press y or else press n.

- **Verifying and setting file permissions**

Listing 31: Verifying and setting file permissions

```
Verifying and setting file permissions
--------------------------------------

Did you install this version with >pkgadd< or did you already verify
and set the file permissions of your distribution (enter: y) (y/n) [y] >>
```

Hit <RETURN>.

- **Configure hostname resolving method**

Listing 32: Configure hostname resolving method

```

Select default Grid Engine hostname resolving method
----------------------------------------------------

Are all hosts of your cluster in one DNS domain? If this is
the case the hostnames

  >hostA< and >hostA.foo.com<

would be treated as equal, because the DNS domain name >foo.com<
is ignored when comparing hostnames.

Are all hosts of your cluster in a single DNS domain (y/n) [y] >>
```

Hit <RETURN>.

- **Configure Grid Engine JMX MBean server**

Listing 33: Configure Grid Engine JMX MBean server

```

Grid Engine JMX MBean server
----------------------------

In order to use the SGE Inspect or the Service Domain Manager (SDM)
SGE adapter you need to configure a JMX server in qmaster. Qmaster
will then load a Java Virtual Machine through a shared library.
NOTE: Java 1.5 or later is required for the JMX MBean server.

Do you want to enable the JMX MBean server (y/n) [y] >>
```

Press n and hit <RETURN>.

- **Configure spooling**

Listing 34: Configure spooling

```

Setup spooling
--------------
Your SGE binaries are compiled to link the spooling libraries
during runtime (dynamically). So you can choose between Berkeley DB
spooling and Classic spooling method.
Please choose a spooling method (berkeleydb|classic) [berkeleydb] >>
```

Enter berkeleydb and hit <RETURN>.

- **Configure Grid Engine group id range**

Listing 35: Configure Grid Engine group id range

```
Grid Engine group id range
--------------------------

When jobs are started under the control of Grid Engine an additional group id
is set on platforms which do not support jobs. This is done to provide maximum
control for Grid Engine jobs.

This additional UNIX group id range must be unused group id's in your system.
Each job will be assigned a unique id during the time it is running.
Therefore you need to provide a range of id's which will be assigned
dynamically for jobs.

The range must be big enough to provide enough numbers for the maximum number
of Grid Engine jobs running at a single moment on a single host. E.g. a range
like >20000-20100< means, that Grid Engine will use the group ids from
20000-20100 and provides a range for 100 Grid Engine jobs at the same time
on a single host.

You can change at any time the group id range in your cluster configuration.

Please enter a range [20000-20100] >>
```

Hit <RETURN>.

- **Configure Grid Engine cluster**

Listing 36: Configure Grid Engine cluster

```
Grid Engine cluster configuration
---------------------------------

Please give the basic configuration parameters of your Grid Engine
installation:

  <execd_spool_dir>

The pathname of the spool directory of the execution hosts. User >root<
must have the right to create this directory and to write into it.

Default: [/grid-engine/ge6.2u5/default/spool]


<administrator_mail>

The email address of the administrator to whom problem reports are sent.

It is recommended to configure this parameter. You may use >none<
if you do not wish to receive administrator mail.

Please enter an email address in the form >user@foo.com<.

Default: [none] >>
```

Enter the values and hit <RETURN>.

- **Finishing the installation**

Listing 37: Creating local configuration

```
Creating local configuration
----------------------------
Creating >act_qmaster< file
Adding default complex attributes
Adding default parallel environments (PE)
Adding SGE default usersets
Adding >sge_aliases< path aliases file
Adding >qtask< qtcsh sample default request file
Adding >sge_request< default submit options file
Creating >sgemaster< script
Creating >sgeexecd< script
Creating settings files for >.profile/.cshrc<
```

```
Hit <RETURN> to continue >>
```

Listing 38: Installing qmaster startup script

```
qmaster startup script
----------------------

We can install the startup script that will
start qmaster at machine boot (y/n) [y] >>
```

Listing 39: Starting qmaster daemon

```
Grid Engine qmaster startup
---------------------------

Starting qmaster daemon. Please wait ...

Hit <RETURN> to continue >>
```

- **Adding Grid Engine hosts**

Listing 40: Adding Grid Engine hosts

```

Adding Grid Engine hosts
------------------------

Please now add the list of hosts, where you will later install your execution
daemons. These hosts will be also added as valid submit hosts.

Please enter a blank separated list of your execution hosts. You may
press <RETURN> if the line is getting too long. Once you are finished
simply press <RETURN> without entering a name.

You also may prepare a file with the hostnames of the machines where you plan
to install Grid Engine. This may be convenient if you are installing Grid
Engine on many hosts.

Do you want to use a file which contains the list of hosts (y/n) [n] >>
```

Listing 41: Adding Admin and Submit hosts

```
Adding admin and submit hosts
-----------------------------

Please enter a blank seperated list of hosts.

Stop by entering <RETURN>. You may repeat this step until you are
entering an empty list. You will see messages from Grid Engine
when the hosts are added.

Host(s):
```

Listing 42: Adding shadow hosts

```
If you want to use a shadow host, it is recommended to add this host
to the list of administrative hosts.

If you are not sure, it is also possible to add or remove hosts after the
installation with <qconf -ah hostname> for adding and <qconf -dh hostname>
for removing this host

Attention: This is not the shadow host installation
procedure.
You still have to install the shadow host separately

Do you want to add your shadow host(s) now? (y/n) [y] >>
```

Listing 43: Creating queues and hostgroup

```
Creating the default <all.q> queue and <allhosts> hostgroup
-----------------------------------------------------------

root@master added "@allhosts" to host group list


Hit <RETURN> to continue >>
```

- **Scheduler Tuning**

Listing 44: Scheduler Tuning

```
Scheduler Tuning
----------------

The details on the different options are described in the manual.

Configurations
--------------
1) Normal
          Fixed interval scheduling, report limited scheduling information,
          actual + assumed load

2) High
          Fixed interval scheduling, report limited scheduling information,
          actual load

3) Max
          Immediate Scheduling, report no scheduling information,
          actual load

Enter the number of your preferred configuration and hit <RETURN>!
Default configuration is [1] >>
```

- **Setting up of environment variables**

Listing 45: Setting up of environment variables

```
Using Grid Engine
-----------------

You should now enter the command:

  source /grid-engine/ge6.2u5/default/common/settings.csh

if you are a csh/tcsh user or

  # . /grid-engine/ge6.2u5/default/common/settings.sh

if you are a sh/ksh user.

This will set or expand the following environment variables:

  - $SGE_ROOT         (always necessary)
  - $SGE_CELL         (if you are using a cell other than >default<)
  - $SGE_CLUSTER_NAME (always necessary)
  - $SGE_QMASTER_PORT (if you haven't added the service >sge_qmaster<)
  - $SGE_EXECD_PORT   (if you haven't added the service >sge_execd<)
  - $PATH/$path       (to find the Grid Engine binaries)
  - $MANPATH          (to access the manual pages)

Hit <RETURN> to see where Grid Engine logs messages >>
```

Add the following line in /etc/bashrc
sh /grid-engine/ge6.2u5/default/common/settings.sh.

- **Viewing the Installation messages**

Listing 46: Viewing the Installation messages

```
Grid Engine messages
--------------------

Grid Engine messages can be found at:

  /tmp/qmaster_messages (during qmaster startup)
  /tmp/execd_messages   (during execution daemon startup)

After startup the daemons log their messages in their spool directories.

  Qmaster:     /grid-engine/ge6.2u5/default/spool/qmaster/messages
  Exec daemon: <execd_spool_dir>/<hostname>/messages


Grid Engine startup scripts
---------------------------

Grid Engine startup scripts can be found at:

/grid-engine/ge6.2u5/default/common/sgemaster (qmaster)
  /grid-engine/ge6.2u5/default/common/sgeexecd (execd)
```

- **Installation complete**

Listing 47: Installation complete

```

Your Grid Engine qmaster installation is now completed
------------------------------------------------------

Please now login to all hosts where you want to run an execution daemon
and start the execution host installation procedure.

If you want to run an execution daemon on this host, please do not forget
to make the execution host installation in this host as well.

All execution hosts must be administrative hosts during the installation.
All hosts which you added to the list of administrative hosts during this
installation procedure can now be installed.

You may verify your administrative hosts with the command

  # qconf -sh

and you may add new administrative hosts with the command

  # qconf -ah <hostname>

Please hit <RETURN> >>
```

3.3.2 Execution Host

Assume the SGE_ROOT directory is /grid-engine/ge6.2u5. The following steps are to be carried out on all the execution hosts.

- **Go to SGE_ROOT directory**

Listing 48: Go to SGE_ROOT directory

```
[root@node1 ~]# cd /grid-engine/ge6.2u5/
```

```
[root@node1 ge6.2u5]#
```

- **Run install_execd**

Listing 49: Run install_execd

```
[root@node1 ~]# ./install_qmaster
```

- **Beginning the installation**

Listing 50: Beginning the installation

```

Welcome to the Grid Engine execution host installation
------------------------------------------------------

If you haven't installed the Grid Engine qmaster host yet, you must execute
this step (with >install_qmaster<) prior the execution host installation.

For a sucessfull installation you need a running Grid Engine qmaster. It is
also neccesary that this host is an administrative host.

You can verify your current list of administrative hosts with
the command:

  # qconf -sh

You can add an administrative host with the command:

  # qconf -ah <hostname>

The execution host installation will take approximately 5 minutes.

Hit <RETURN> to continue >>
```

- **Checking $SGE_ROOT directory**

Listing 51: Checking $SGE_ROOT directory

```
Checking $SGE_ROOT directory
----------------------------

The Grid Engine root directory is:

  $SGE_ROOT = /grid-engine/ge6.2u5

If this directory is not correct (e.g. it may contain an automounter
prefix) enter the correct path to this directory or hit <RETURN>
to use default [/grid-engine/ge6.2u5] >>
```

Hit <RETURN>.

- **Configure Grid Engine Cells**

Listing 52: Configure Grid Engine cells

```
Grid Engine cells
 -----------------

 Please enter cell name which you used for the qmaster
 installation or press <RETURN> to use [default] >>
```

- **Configure Grid Engine TCP/IP communication service**

Listing 53: Configure Grid Engine TCP/IP communication service

```
Grid Engine TCP/IP communication service
 ----------------------------------------
```

```
The port for sge_execd is currently set by the shell environment.

  SGE_EXECD_PORT = 6445

Hit <RETURN> to continue >>
```

- **Checking hostname resolving**

Listing 54: Checking hostname resolving

```
Checking hostname resolving
 ---------------------------

 This hostname is known at qmaster as an administrative host.

 Hit <RETURN> to continue >>
```

- **Configure Grid Engine execd spool directory**

Listing 55: Configure Grid Engine execd spool directory

```
Execd spool directory configuration
 -----------------------------------

 You defined a global spool directory when you installed the master host.
 You can use that directory for spooling jobs from this execution host
 or you can define a different spool directory for this execution host.

 ATTENTION: For most operating systems, the spool directory does not have to
 be located on a local disk. The spool directory can be located on a
 network-accessible drive. However, using a local spool directory provides
 better performance.

 FOR WINDOWS USERS: On Windows systems, the spool directory MUST be located
 on a local disk. If you install an execution daemon on a Windows system
 without a local spool directory, the execution host is unusable.

 The spool directory is currently set to:
 <</grid-engine/ge6.2u5/default/spool/node1>>

 Do you want to configure a different spool directory
 for this host (y/n) [n] >>
```

- **Finishing the installation**

Listing 56: Creating local configuration

```
Creating local configuration
----------------------------
root@node1 modified "node1" in configuration list
Local configuration for host >node1< created.

Hit <RETURN> to continue >>
```

Listing 57: Installing execd startup script

```
execd startup script
--------------------

We can install the startup script that will
start execd at machine boot (y/n) [y] >>
```

Listing 58: Starting execd daemon

```
Grid Engine execution daemon startup
------------------------------------
```

```
Starting execution daemon. Please wait ...
  starting sge_execd

Hit <RETURN> to continue >>
```

Listing 59: Adding queue

```
Adding a queue for this host
----------------------------

We can now add a queue instance for this host:

      - it is added to the >allhosts< hostgroup
  - the queue provides 8 slot(s) for jobs in all queues
    referencing the >allhosts< hostgroup

You do not need to add this host now, but before running jobs on this host
it must be added to at least one queue.

Do you want to add a default queue instance for this host (y/n) [y] >>

root@node1 modified "@allhosts" in host group list
root@node1 modified "all.q" in cluster queue list
```

- **Setting up of environment variables**

Listing 60: Setting up of environment variables

```

Using Grid Engine
-----------------

You should now enter the command:

  source /grid-engine/ge6.2u5/default/common/settings.csh

if you are a csh/tcsh user or

  # . /grid-engine/ge6.2u5/default/common/settings.sh

if you are a sh/ksh user.

This will set or expand the following environment variables:

  - $SGE_ROOT         (always necessary)
  - $SGE_CELL         (if you are using a cell other than >default<)
  - $SGE_CLUSTER_NAME (always necessary)
  - $SGE_QMASTER_PORT (if you haven't added the service >sge_qmaster<)
  - $SGE_EXECD_PORT   (if you haven't added the service >sge_execd<)
  - $PATH/$path       (to find the Grid Engine binaries)
  - $MANPATH          (to access the manual pages)

Hit <RETURN> to see where Grid Engine logs messages >>
```

Add the following line in `/etc/bashrc`
`sh /grid-engine/ge6.2u5/default/common/settings.sh`.

- **Viewing the Installation messages**

Listing 61: Viewing the Installation messages

```

Grid Engine messages
--------------------

Grid Engine messages can be found at:

  /tmp/qmaster_messages (during qmaster startup)
  /tmp/execd_messages   (during execution daemon startup)
```

```
After startup the daemons log their messages in their spool directories.

  Qmaster:     /grid-engine/ge6.2u5/default/spool/qmaster/messages
  Exec daemon: <execd_spool_dir>/<hostname>/messages


Grid Engine startup scripts
---------------------------

Grid Engine startup scripts can be found at:

  /grid-engine/ge6.2u5/default/common/sgemaster (qmaster)
/grid-engine/ge6.2u5/default/common/sgeexecd (execd)
```

4 Using Oracle Grid Engine

The cluster can be used in two modes - **Interactive Mode** and **Batch Mode**. All the commands mentioned are to be executed on the master host.

- **Interactive Mode**
  The interactive mode is intended for getting a session on one of the execution hosts. It is like working on the execution host as on a usual PC). This is useful for testing if the job works correctly in a cluster node environment.

  After getting a shell, a slot on the Grid Engine queue is reserved, so if someone else submits a large array job, the interactive job will still be assigned a processor.

  Commands list :

  - `qsh` : Starts an interactive X Window session. Finds a free computing node and gives an xterm there. X Window forwarding must be enabled.
  - `qrsh` : Similar to qsh, but console session instead of a graphical one.
  - `qlogin` : Starts an interactive login session.

- **Batch Mode**

  The batch mode is intended for submitting a job to execution hosts where it will be run. Grid engine will automatically start the job on the execution hosts with the least load.

  Commands list :

  - **Submitting a job**

    Listing 62: Submitting a job

    ```
    [user@master ~]# qsub [options] program_name [program_args]
    ```

    Useful Options

    * -q : use a particular task queue, e.g. `-q serial.q@comp00`
    * -l : demand certain resources, e.g. memory, for the job. Sample:
      `-l vf=2G`
      will request 2G of free virtual memory for the job so that it won't be scheduled to a node where this amount of memory is unavailable.
    * -m : mail you when job done
    * -cwd : the files will be put in the working directory rather than your home directory
    * -o : Specify the std_out file, e.g. `-o stdout.txt`
    * -e : Specify the std_err file, e.g. `-e stderr.txt`

  - **Viewing jobs**

Listing 63: Viewing jobs

```
[user@master ~]# qstat              ## Displays jobs of all users
[user@master ~]#
[user@master ~]# qstat -u username  ## Displays jobs of username
[user@master ~]#
```

- **Deleting job**

Listing 64: Deleting job

```
[user@master ~]# qdel [job_id]      ## To delete a particular job
[user@master ~]#
[user@master ~]# qdel -u username   ## To delete all jobs of username
[user@master ~]#
```

- **Generating Accounting statistics**

Listing 65: Generating Accounting statistics

```
[user@master ~]# qacct -j job_id       ##  Get detailed listing of a particular
[user@master ~]#                       ##  job - once it has finished running
```

- **Viewing hosts**

Listing 66: Viewing Hosts

```
[user@master ~]# qstat -f          ##  view all the hosts and their current load
[user@master ~]#
```

5 References

- Beginner's Guide to Oracle Grid Engine 6.2 (An Oracle White Paper, August 2010)
- Linux Home Networking