Scalable Systems

# Integration of SGE into NPACI Rocks

Najib Ninaba
Lead Developer
Scalable Systems Pte Ltd
najib@scalablesys.com
najib@scs.com.sg

# About

Scalable Systems

- Scalable Systems Pte Ltd
  - formerly the HPC team in Singapore Computer Systems Linux Competency Centre
    - http://www.scs.com.sg/lcc
  - Co-developer of NPACI Rocks
  - Maintains various 3rd party RPM packaging effort in NPACI Rocks :
    - GridEngine, PVFS, Myrinet, etc

# Motivations

Scalable Systems

- Started initial packaging to replace OpenPBS in Rocks.
- No other existing RPM packaging effort to be found circa December 2001.
- Released first ever(?) GridEngine RPM on 23 January 2002.

# Initial Versions

Scalable Systems

- First version of GridEngine RPM was a binary only package which contains the linux version of the courtesy binaries (ver. 5.2.3) in January 2002.
- Later versions after that contains proper source tarball from GridEngine CVS. Both RPM and the SRPM were made available.

# Announcing GridEngine RPM

Scalable Systems

- Word got around and was suggested by some, Ron Chen among others, to contact the GridEngine community regarding the availability of GridEngine RPM.
- It was during this time that SuSE, namely Anas Nashif, were known to also package GridEngine for that distro.

# GridEngine RPM Feedback

Scalable Systems

- People started to use the RPM and really interesting bugs on the RPM packaging start coming in.
- With such feedback coming in, more improvements and fixes were made to make the packaging more robust.

# GridEngine RPM Features

Scalable Systems

- Tracks closely to the GridEngine stable releases.
- With permission from Anas Nashif from SuSE, incorporated some of his patches also.
- Two variants of the GridEngine RPM package exist. One for generic Red Hat package, the other for NPACI Rocks.

# GridEngine RPM Features

Scalable Systems

- Both variants have the following things in common:
  - Installs under /opt/sge.
  - Creates an entry for sge_commd under /etc/services.
  - Creates GridEngine Admin user for running the GridEngine daemons.
  - When installed, just need to start the rcsge service and that node will be the qmaster/schedd. Other nodes just need to run the install_execd script to join the GridEngine pool (thru NFS).

# Generic Red Hat Version

Scalable Systems

- The generic Red Hat variant were also used as the base by other distros, most noticeably: MSC.Linux and WareWulf cluster project.
- This variant is more commonly used for a typical GridEngine cluster setup that relies on NFS.

# NPACI Rocks Version

Scalable Systems

- The NPACI Rocks variant was customised to integrate GridEngine very closely with Rocks.
- The end goal is to have GridEngine automatically installed and running in Rocks cluster with minimum fuss, ready to schedule jobs.

# NPACI Rocks Version

Scalable Systems

- This variant do not rely on NFS, each node locally installs own binaries, configuration and spool directories for scalability reasons.
- Some rational defaults were setup for GridEngine out of the box.
- Parallel Enviroments for MPI/MPICH also were setup out of the box.

# GridEngine Rocks Status

Scalable Systems

- With such customisations and close integration of Rocks and GridEngine, it was packaged as part of base Rocks version 2.3.0 and later.
- Not yet the default scheduler (OpenPBS currently is) in Rocks but a simple switch will enable GridEngine:

**# touch /etc/USESGE**

**# source /etc/profile.d/gridengine.sh**

# GridEngine Rocks Out Of The Box

Scalable Systems

- With every installation of Rocks now contains a functional GridEngine setup. **Zero Hand Configuration.**
- Rocks admin do not need to do anything other than turning on GridEngine (/etc/USESGE).
- The Rocks infrastructure automatically installs default queues and runs the execd daemons on the compute nodes.

# Future GridEngine & Rocks

Scalable Systems

- Targeting to make GridEngine as the default scheduler by next version (currently Rocks is version 3.0.0)
- There is current effort to setup Globus with Rocks and GridEngine automatically.
- Also, currently intending to add RPM support to mk_dist script so that RPMs can be generated from a CVS tarball of GridEngine.

# Resources

- Http://rocksclusters.org/
- http://www.scalablesys.com/

# The End