Sun Grid Engine at OSC

Fred Youhanaie
fy@comlab.ox.ac.uk
Oxford Supercomputing Centre
http://www.osc.ox.ac.uk/

Overview

- History of OSC
- Current Environment: H/W and SGE Configuration
- The local accounting database
- Integration with Globus
- Concluding remarks, and wish list :-)

History

- 1997 OSC Established
  - A collaboration between about 20 research groups in Oxford University's Biochemistry, Physics, Physiology, Computer Science, Materials, Engineering Sciences, Earth Sciences, Physics and Chemistry Departments.
- 1998: First system arrived
  - SGI/Cray Origin 2000, 72-96 CPUs, IRIX, Miser
  - Shared Memory, MPI, PVM, BSP
  - In short: self contained monolithic system
- 2002: SGI replaced with two clusters from Sun and IBM
  - Shared Memory System - Sun/Solaris
  - Distributed Memory System - IBM/Linux
- There are currently 29 groups with a total of 140 users

Shared Memory System

- 4 x Sun Fire 6800, UltraSPARC III 900MHz, 2GB/CPU
- 84 CPU in total, 24+12+24+24
- Sun HPC CT 5.0
- OpenMP, MPI, BSP, adhoc multithreaded
- Used as four individual machines
- Two Sun V880s used as file server and head node
- 1TB disk storage + backup facility

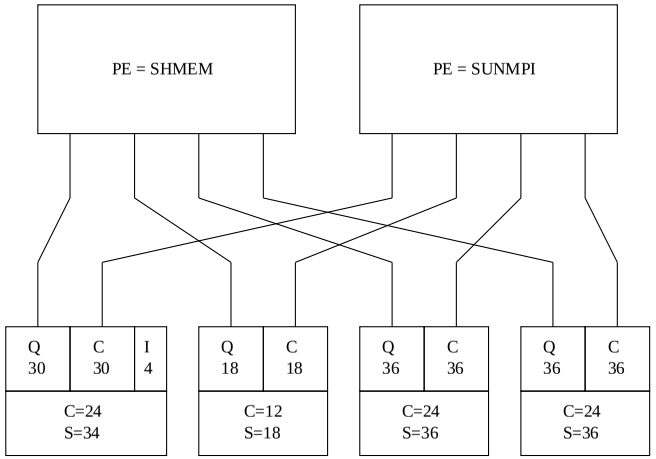
Distributed Memory System

- 64 x IBM x330 dual CPU, 1.26GHz Pentium III, 2GB/node
- Myrinet 2000 switch
- Management nodes and File server (x232, x342 and x330)
- RedHat 7.2, kernel 2.4
- Initially PBS, now under SGEEE
- MPICH-GM, MPICH-P4, PVM, BSP

SGE Basic Configuration

- SGE first installed on the Sun cluster in March, 2002.
- IBM/Linux cluster converted from PBS to SGE in Sep., 2003.
- All system under one CELL.
- Currently running SGEEE 5.3p4 (locally compiled)
- Users are not allowed to select queues/hosts
- Access via Parallel Environment only
- No BATCH queues - use `'-pe <PE> 1'` for serial jobs
- Users select platform by selecting the appropriate PE
  - Except for interactive jobs!

SGE Config. - Sun/Solaris

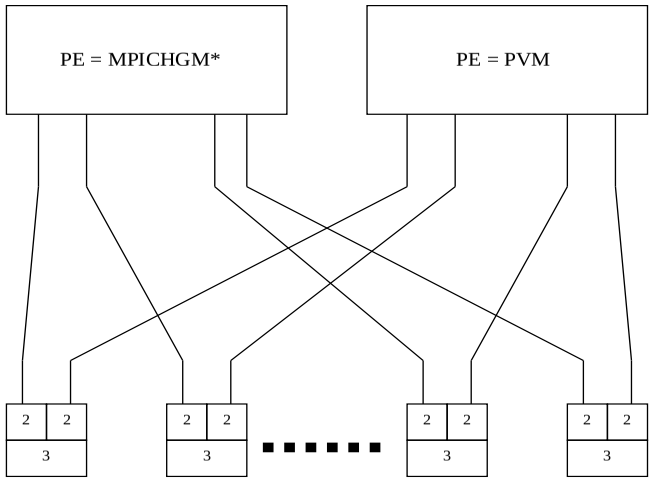


4 hosts, two queues per host, $slots = 1.5 \times CPUs + 1$ interactive Q

SGE Config. - Sun/Solaris (2)

- 1 interactive queue
  - uses calendar - Mon-Fri 8am-8pm
  - uses processor set, created/destroyed via cron
  - $h\_rt \leq 1hour$
  - Minimizes idle CPU time.
- PE: shmem
  - 4 x general purpose (OpenMP, BSP)
  - no `pe_start` or `pe_stop` script
  - mainly used for OpenMP jobs
- PE: sunmpi
  - 4 x CRE - for HPC CT - Sun MPI - SHM only
  - Tight integration for HPC ClusterTools

SGE - IBM/Linux



64 nodes, 2 queues per node, 3 slots per host (but only 2 CPUs)

SGE - IBM/Linux (2)

- 64 nodes, 2 queues per node, 3 slots per host
  - short queue ($h\_rt \leq 10min$) - 2 slots
    - coarse grain interactive
  - long queue ($h\_rt \leq 504hours$) - 2 slots
- PEs: mpichgm1, mpichgm2, mpichgm
- PEs: pvm - $allocation\_rule = 2$
  - PVM does not like more than one job per user per host
  - Only used for one application
- A recent change
  - Top 8 queues/nodes only available to 2 slot PEs, i.e. PVM and mpichgm2
  - Bottom 8 queues/nodes only availale to 1 slot PEs, i.e. mpichgm and mpichgm1

Share Tree

- The 29 groups entitled to equal share of the resources
- We are only concerned with CPU time
- 3 levels: Root - Group - Users with the group
- Not all group leaders have opted for subdivisions
- For each group we have:
  - Access list - userset
  - Project
  - Projects restricted to group with same name
- Each user:
  - Belongs to one group only
  - Has a default project

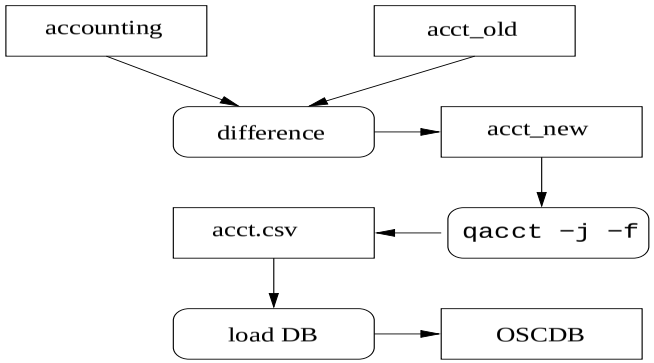
Share Tree (2)

- Before Grid Engine (SGI/Miser)
  - Usage was tracked with locally developed software
  - Usage was checked automatically
  - Those above their quota were prevented from submitting jobs
  - Even though there were free CPUs available
- SGE Share Tree
  - No need to keep track of usage externally
  - But, no means of preempting/stopping overdrawn users
  - Overloading helps

The OSC Database

- `PostgreSQL`
- accounting data from SGE
- downtime figures
- sge_sharemon data
- miscellaneous sundry table
- performance data from sar
- Acounting data most important as we need to provide regular usage reports.

Accounting System



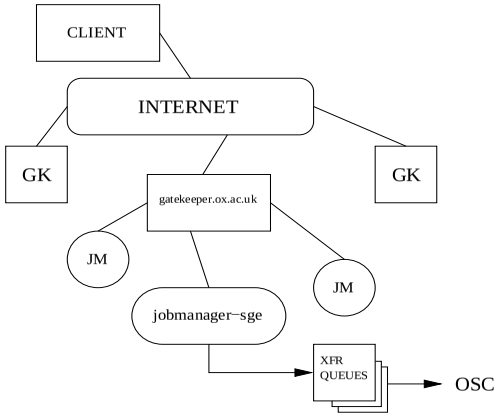
- cron job runs every 20 minutes
- `Category` field not stored in database
- This is a single table `sgeacct`

Accounting System (2)

- But, `sgeacct` not always convenient.
- Tight integration produces multiple records per job.
- Hence view `jobacct`
- Some fields copied across: *gname*, *owner*, *project*, *department*, *jobname*, *jobnumber*, *account*, *priority*, *qsub_time*, *granted_pe*, *slots*.
- Other fields aggregated:
  - $start\_time = min(start\_time)$ i.e. the earliest start_time.
  - $end\_time = max(end\_time)$ i.e. the latest end_time.
  - $cpu = sum(cpu)$
  - $wallclock = max(end\_time) - min(start\_time)$

Globus Integration (GT2)

```
globus-job-run gatekeeper.ox.ac.uk/jobmanager-sge \
               -q qid prog args
```



Example above does not show security infrastructure

Globus Integration (GT2)

- jobmanager-sge from Imperial College - Marko Krznaric
- gatekeeper runs qmaster and execd with a single local queue
- One local queue for local execution - more orderly that jobmanager-fork
- transfer queues - one per OSC parallel environment
- Transfer queues hide OSC details
- They also allow for multiple GE clusters through a single jobmanager
- Currently using NFS for file staging
- We will probably use parts of ToG for file staging

Possible Enhancements

- Queue based usage scaling
  - Higher premium for interactive queues
  - Lower usage scaling for some queues.
  - Will involves source code modification.
- May try the Maui plugin.

Concluding Remarks

- Very happy with the mailing list response :-))
- Many ways of configuring the system.
- Wishlist (probably covered in Release 6.0)
  - Preemptive scheduling
  - Low ticket small jobs overtake high ticket large jobs
  - Accounting system needs a better structure
    - Include records for deleted jobs
    - Master/detail records per job run

Acknowledgement

- Sun Microsystems
- Colleagues at OSC
  - Bob McLatchie
  - Joe Pitt-Francis
  - Jon Lockley