HPC Scheduling

A summary of scheduler use across sites
Leeds workshop (Sept 2016)

Christopher J. Walker
C.J.Walker@qmul.ac.uk

Schedulers Used

- Schedulers
  - Gridengine (SGE)
    - Sun of Grid engine (SoGE)
    - Open Grid Scheduler (OGS)
    - Univa GridEngine (UGE) - Commercial
  - Loadleveller
  - LSF
  - PBS
  - Slurm
  - Torque + Moab/Maui

Notes

- More than one scheduler at a site
  - Different clusters
    - Same or different team
  - Migration
- Data collected as a “Who to talk to about scheduler X”
- Slides updated and corrected before upload
- https://twiki.cern.ch/twiki/bin/view/LCG/BatchSystemComparison

Gridengine derivatives

- Edinburgh (OGS)
- Hartree (also LSF, SGE, Slurm)
- Lancaster (SoGE)
- Leeds (SoGE)
- LSE (SGE) – consisdering Slurm/moab
- Manchester (Gridengine)
- QMUL (SGE/UGE), (GridPP: Slurm)
- QUB: Torque/Maui → SGE
- Strathclyde (OGS)
- Sussex (UGE)
- UCL (SoGE, also PBS pro)
    - Thomas *EPSRC Tier-2* (SoGE)

Loadleveller,LSF, PBS

- Loadleveller
  - STFC (Daresbury) (Also LSC, Slurm, SGE)
- LSF
  - STFC (see above)
  - UEA
  - Hartree (Also LSF, SGE, Slurm)
- PBS
  - Cranfield
  - Cirrus (EPSRC Tier-2)
  - GW4 (EPSRC Tier-2)
  - Imperial

Slurm

- Bath
- Bristol (Torque/Moab older systems)
- Crick
- Hartree (also LSF, SGE)
- Hull
- Loughborough
- QMUL-GridPP (ITSR SGE->UGE)
- STFC-Daresbury (also LSF, loadleveller,sge)
- Warwick (Slurm/Slurm, Moab/Slurm, Moab/Torque)
- Jade - *EPSRC Tier-2*
- HPC Midlands+ - *EPSRC Tier-2*

Torque + Maui/Moab

- Birmingham (Torque/Moab)
- Bristol (+ Slurm on newer systems)
- EPCC (Torque/Maui, pbspro (free))
- Leicester (Torque/Moab)
- QUB (Torque/Maui and SGE)
- Southampton (Torque/maui , Torque/Moab (newer))

Summary

- Lots of different schedulers in use
- Lots of people to ask for advice
- EPSRC Tier-2 schedulers added

Scheduler talks

- Slurm (Steve Chapman)
  - + New, free (but can pay for support)
  - + Users and sysadmins happy
  - + Rich config options and good docs
  - - Large initial effort
  - - Lacked allocation and finance tracking

Cloud Matt Harvey (m.j.harvey@imperial.ac.uk)

- cloud capex -> opex
  - Cost clarity
- respond to load spikes
- Possible hybrid cluster
- Note RCUK cloud working group

CJW cloud comments

- Scheduling on the cloud
  - Infinite cloud?
  - Kubernetes, swarm etc ?

Mixed scheduling

- Should depts buy their own nodes
- Mixed node types makes scheduling more difficult
  - Notably GPUs
- Users
  - May game system
  - May not understand system

Services and data migration issues

- IO a challenge
  - Burst buffers ?
    - Slurm has support
  - Stage in/out?
    - Most sites don't
  - Data locality
- Scratch not scratchy
- Small files inefficient
- Cloudbursting
  - CJW comment – do we need a mentality change for cloud?

Cost

- Procurement
  - Scheduler cost often hidden
    - (and a surprise on renewal)
  - What will procurement let you buy
- Open source or proprietary
  - Functionality
  - Support contract for open source?
    - Quicker fixes going via github patches than via formal support route for torque
  - Service credits used to improve open source?

Conclusions

- Lots of useful discussions
  - Lots of knowledge in the community
- IO as always a question
  - data locality (particularly cloud)
  - Small files
- Ash may come up with a more detailed summary