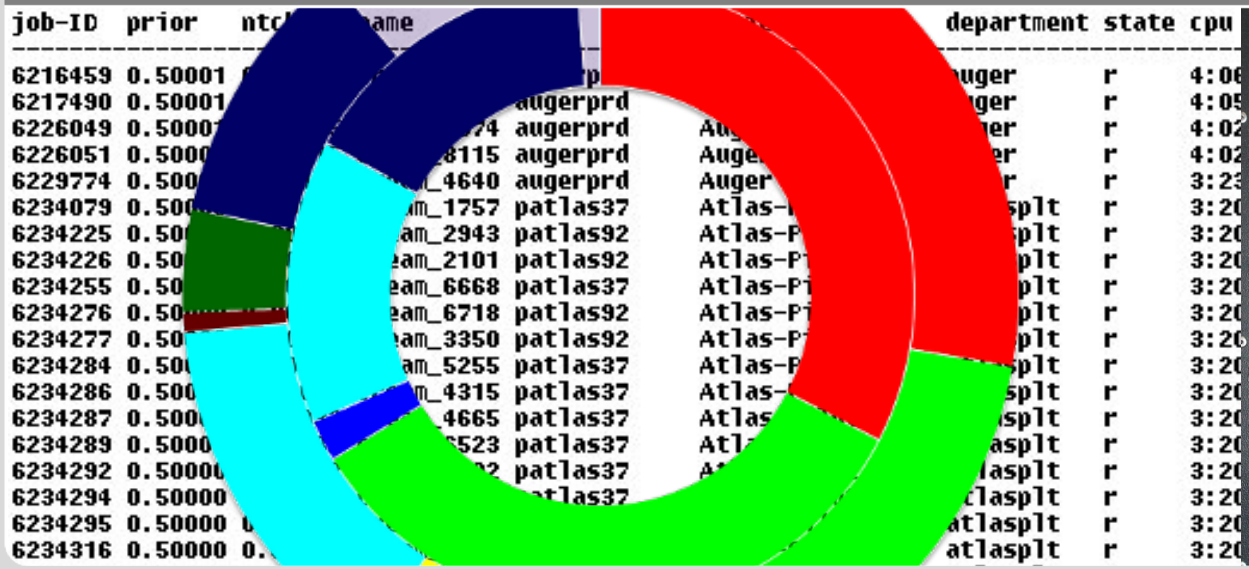
Karlsruhe Institute of Technology

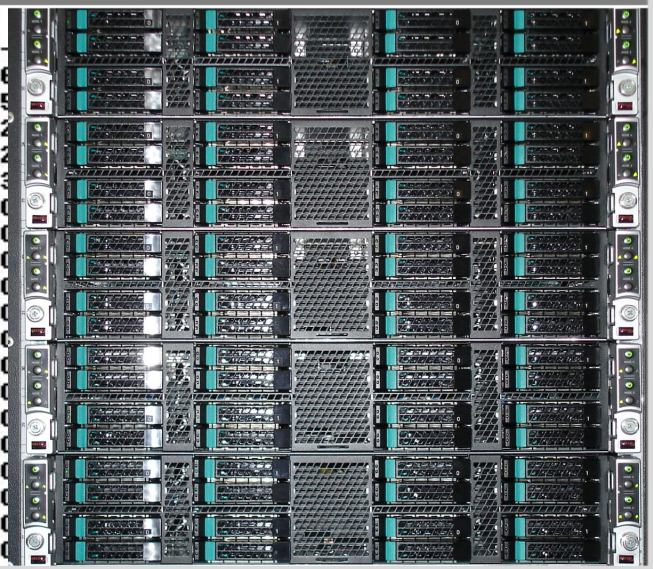
SCC
Steinbuch Centre for Computing

GridKa Site Report
HEPiX Spring 2013

Manfred Alef

STEINBUCH CENTRE FOR COMPUTING – SCC





LRMS Migration

LRMS Migration

- GridKa cluster:
  - Single cluster
    - Fair-share policy configured according to pledged VO shares
  - 146 kHS06
  - 1,000 worker nodes
  - 11,800 (physical) cores
  - 17,400 logical (hyperthreaded) cores
  - **14,600 job slots**
  - **~ 1,500,000 jobs per month**

LRMS Migration

- LRMS used until 2012:
  - PBS Professional
    - Frequent troubles, e.g.
      - Black hole nodes
      - Corrupted job descriptions in database
    - More and more issues with every cluster expansion
    - Bug in new release of licence management tool caused downtime of 3+ days
    - 2010: Split into 2 sub-clusters to improve stability
      - More stable than before, but still frequent trouble

LRMS Migration

- Looking for alternative LRMS (2010 ... 2011):
  - Several LRMS' tested
  - SGE+UGE test installations on whole farm (in parallel with PBS)
    - Up to 1000 job slots per WN configured
    - Hammered with Millions of jobs (sleepers)
- Small test cluster (~ 2,000 slots) managed by UGE since summer 2012

LRMS Migration

- Dezember 2012: migration of the whole cluster to UGE
  - Testing phase till March 2013
  - No serious problems so far
- April 2013: UGE licence agreement extended for 2 years

LRMS Migration

- Some configuration details:
  - Single server, no failover
  - Flat files (no DB)
  - Certificate Security Protocol (CSP) enabled
  - Fair-share configurations:
    - Based on reserved usage (aka walltime)
    - Share-tree (using historical data) and functional ticket policy, weighting: 50% each

LRMS Migration

- Experiences:
  - Learning curve
    - Documentations sometimes not clear, or incomplete
    - Some bugs found (uncritical)
    - Very quick and efficient response by Univa support desk
      (located in Germany; no time shift – same time zone as Karlsruhe)
  - Stable operation, no crashes, no black hole node issues
  - Smooth fair-share scheduling, no complains by VOs / users

LRMS Migration

- Missing important feature:
  - Monitoring of job efficiencies (cpu-per-walltime ratio of running jobs)
    - Essential for site and users VOs
    - CPU usage reported by (original) qstat is reserved CPU usage (aka walltime), according to fair-share settings
    - Has been discussed with developper – Univa has already provided a patch to provide consumed CPU as well as walltime

```
# qstat -j \* | grep ^usage | head -5 | cut -d, -f-2
usage    1:                 wallclock=4:11:08:30, cpu=4:08:59:00
usage    1:                 wallclock=4:10:54:56, cpu=4:10:25:41
usage    1:                 wallclock=4:10:19:25, cpu=4:10:02:35
usage    1:                 wallclock=4:07:06:39, cpu=00:00:09
usage    1:                 wallclock=4:07:05:33, cpu=4:04:57:06
#
```

Frequent WN Hard Disk Failures

Frequent WN Hard Disk Failures

- WNs ordered in October 2011
- ~1 host (of 236) crashed per day
  - smartctl reported either 1 or 65535 pending sectors
- Vendor applied firmware patches:
  made worse, now 3 – 4 hosts per day affected
- Vendor replaced all disks:
  problem disappeared
- Supposing Thai flood issue, and / or vibrations
  (fans possibly speedup to multiple of disk rotating speed?)

Questions, Comments?