Oxford Particle Physics
HepSysMan 2008

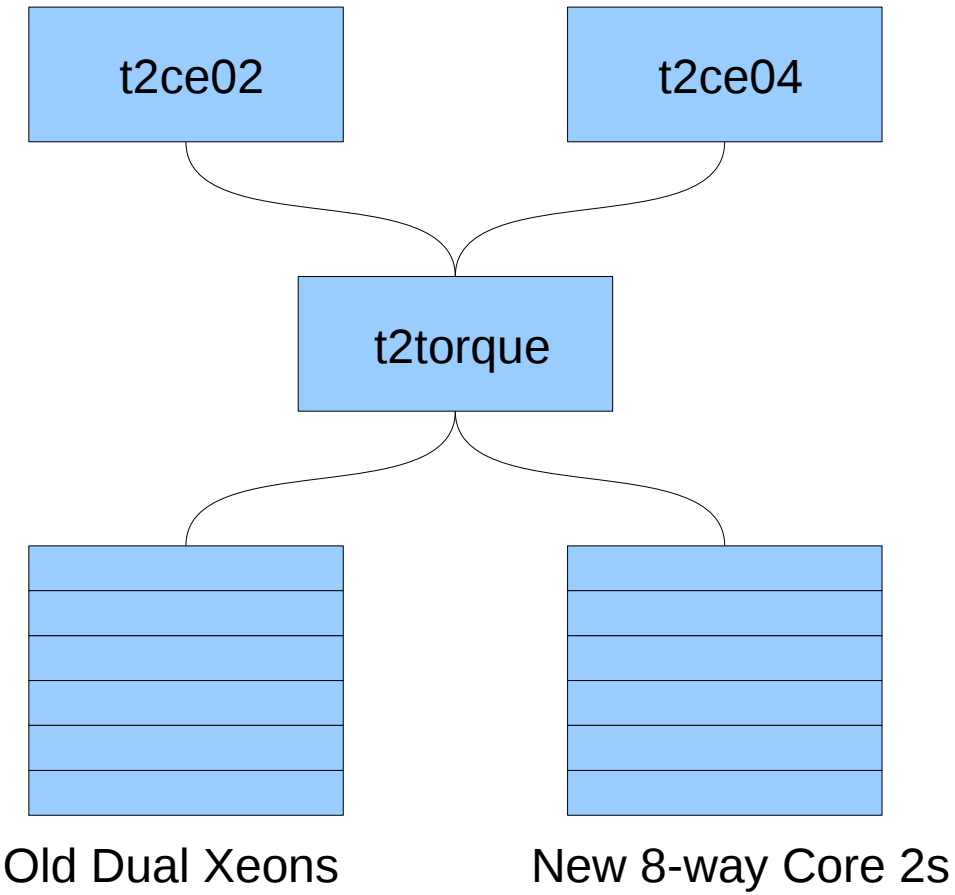
Site Report

Grid cluster: UKI-SOUTHGRID-OX-HEP

- New hardware bought in September 07, 176 CPU cores, ~100Tb of storage.

- 11 'twin' units
  - 22 WNs,
  - 2GB per core
- 11 disk servers
  - 14 750Gb disks
  - RAID 6
  - Plus RAID pair of 250Gb OS disks

Dual CE/Subcluster setup



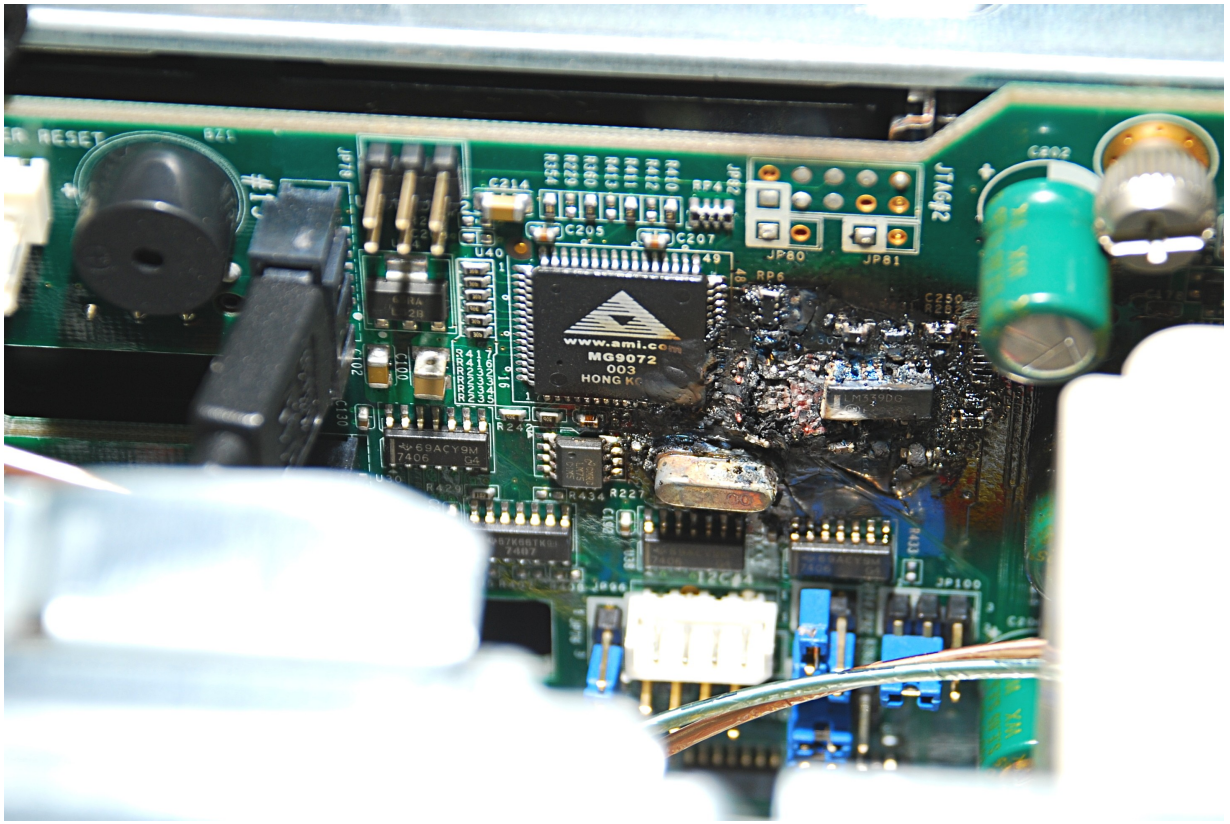
- The system has seven queues:
  - shortdual
  - mediumdual
  - longdual
  - shortoct
  - mediumoct
  - longoct
  - express
- Each CE advertises appropriate queues and properties.
- Express jobs can run anywhere.

Grid cluster - Interesting bits

- We use virtual machines for several service nodes:
  - All CEs, torque server, monitoring server, BDII.
  - The MON will be but isn't yet.
  - The SE front end isn't and probably won't be.
- The SE front end was recently upgraded from an SL3 system to an SL4 system on new hardware.
  - We had a preparatory saga with dpm-drain
  - But the actual upgrade went remarkably smoothly.
  - Except for an apparantly wide-spread and harmless information system bug.
- We moved all the kit to our new room at Begbroke Science Park.
  - By doing it in stages we managed it with only a day and a half's site downtime.
- The MON, BDII and (soon to die) t2ce03 are the only remaining SL3 nodes.

Smokin'

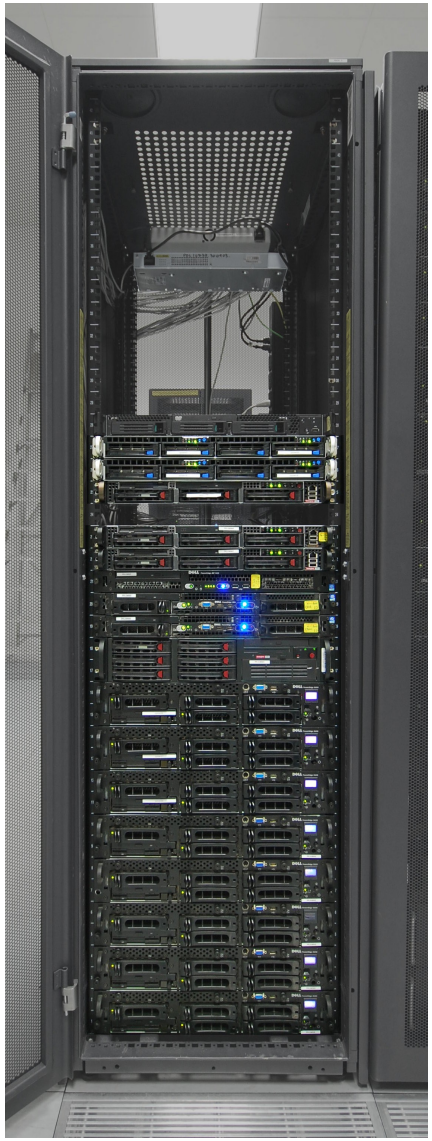
- We had one machine exhibit the (now) fairly well known Supermicro exploding disk server bug.
  - Affected part was replaced pretty quickly, others still pending a preventative replacement.



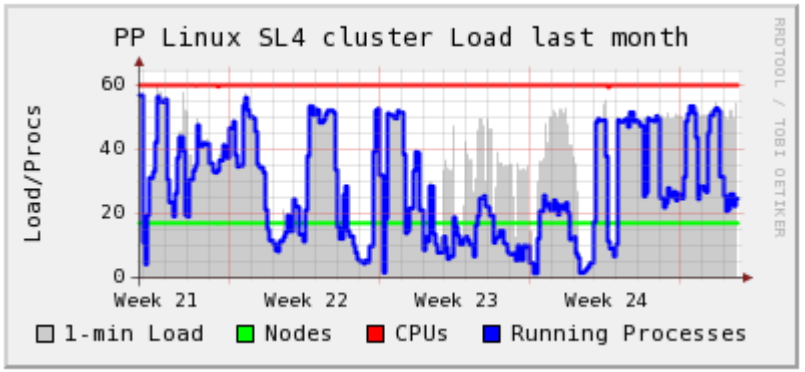
Local cluster storage

- We have three data fileservers:
  - pplxfs2: ~20Tb over three Infortrend SCSI arrays (RAID5)
  - pplxfs3: ~8Tb as a single 3ware internal RAID (RAID6)
  - pplxfs4: ~20Tb as a single 3ware internal RAID (RAID6)
- All are running ext3 over LVM, but (clearly) we can't manage storage between servers.
- We're considering clustering options - GFS, Lustre, OCFS, others?

Local cluster worker nodes

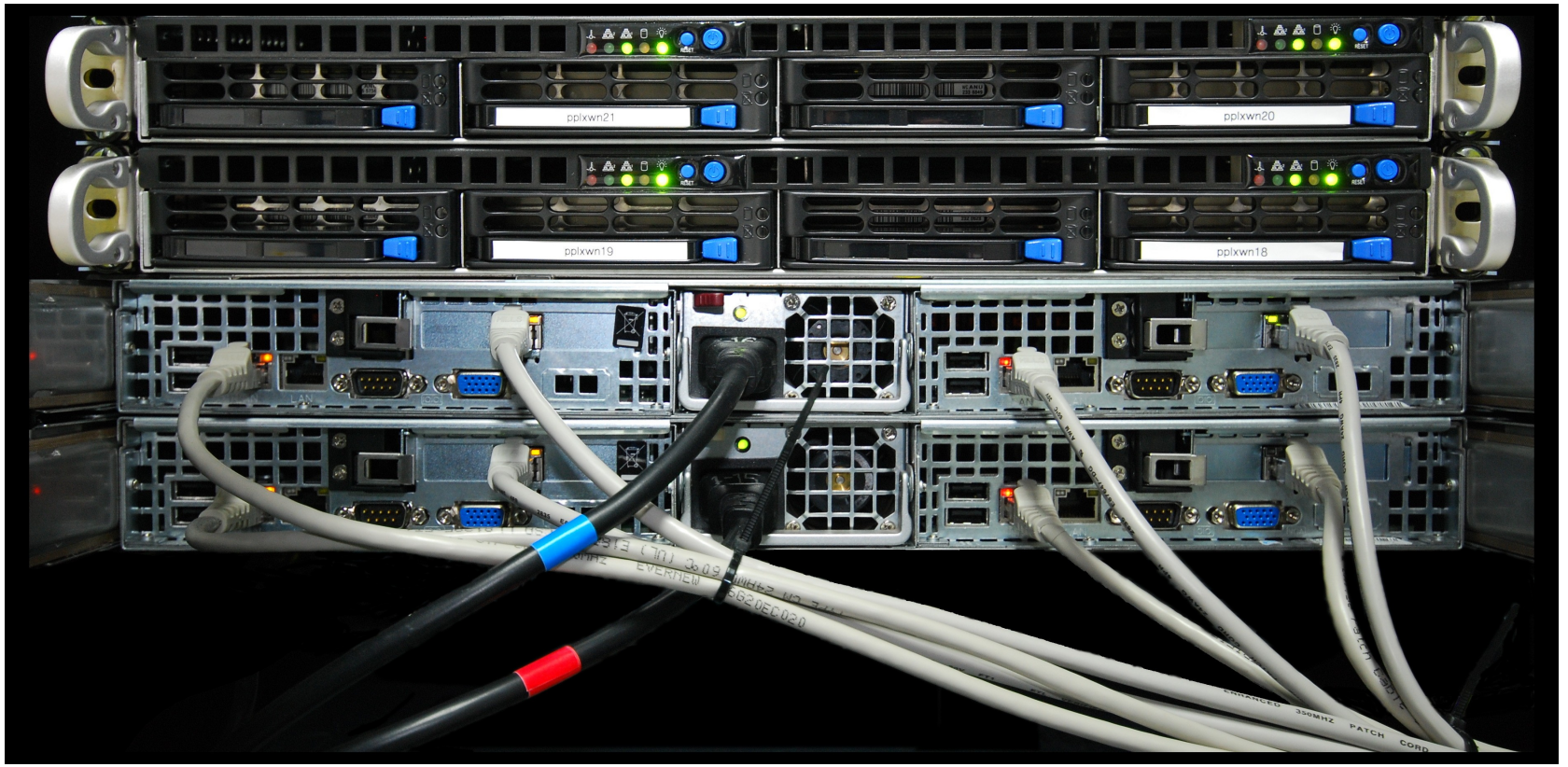


- The local cluster has a total of 60 processors
  - 32 are in a pair of 'twin' worker nodes
  - the rest are old dual processor nodes
- All runs a dead simple setup with NIS user accounts, NFS mounted home and data areas.
- Configured with kickstart, cfengine and some small scripts. We can nuke a node and having back and running in a few minutes.
- Not as consistently busy as it could be.
  - Considering filling in with some kind of grid job.



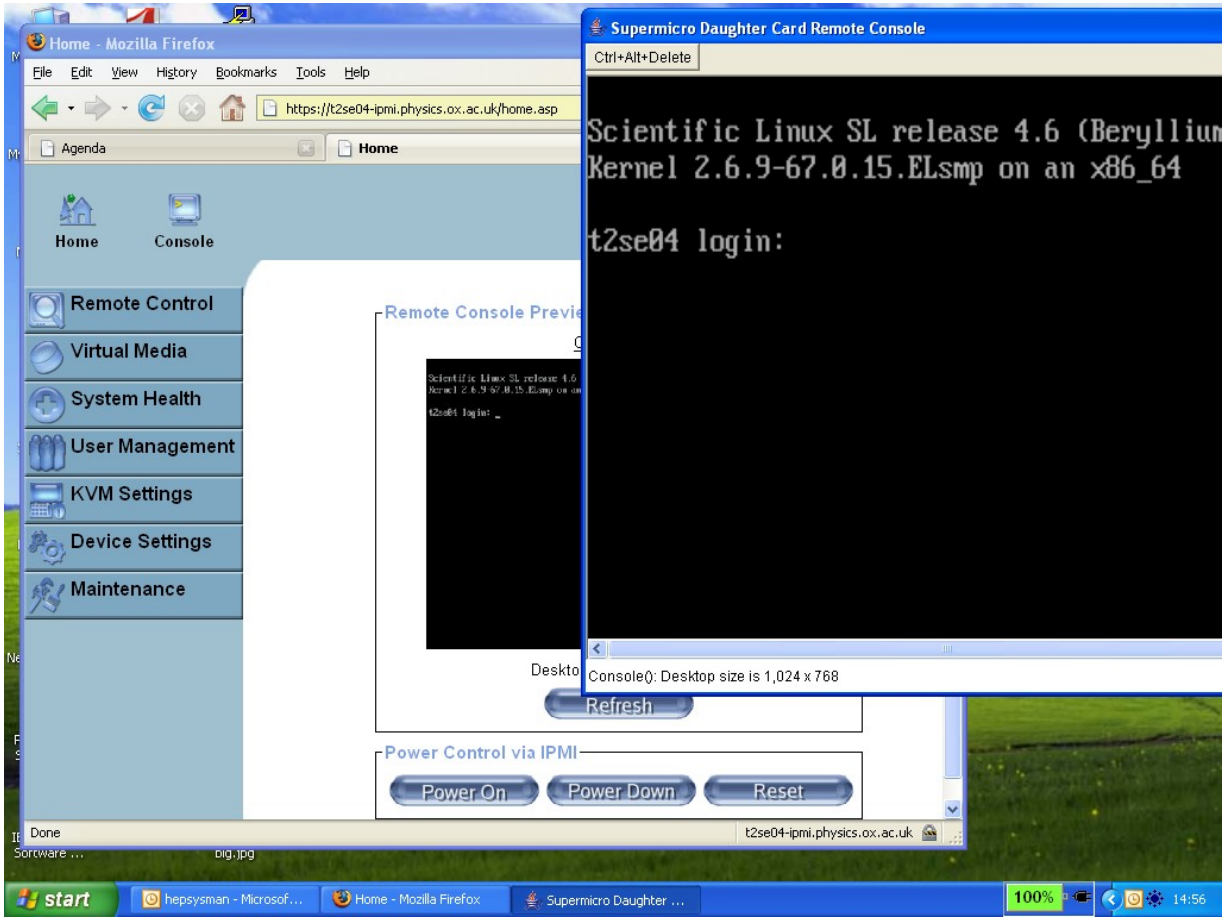
Supermicro 'Twins'

- Each twin has a single PSU feeding two motherboards.
- Ours have the optional KVM over LAN IPMI cards.



IPMI KVM over LAN

- IPMI card allows 'local console' access, including to the BIOS. Really great in principle, mostly good in practice, occasionally painful.



Odds and ends

- We used Amazon's EC2 system to great effect to test scalability of a research group's code.
  - It allowed us to create an image then boot *n* instances of it.
  - We got *n* up to just under a thousand, at a total project cost of about £150
- We had a user with particularly IO heavy jobs running over a static dataset. We rsync-ed the data to some WN's local disks, which worked, but there has to be a better way.
- Extra SL repositories; consider Fedora EPEL as an alternative to Dag; it works for us.
- Nexsan 'SATABeast' Fibrechannel RAID arrays - we don't have any. Do you?