Imperial College
London

Status of GridPP Cloud Sites

Adam Huffman

Agenda

- Imperial status
- Status of other sites
- EGI Federated Cloud
- Usage
- Moving to production
- IPv6

Imperial Cloud Status

- New compute node in service (40 core, 256GB)
  - Attracts all new instances because the scheduler prefers free RAM
  - Benefits of Intel Turbo = ~0.2-0.3 GHz extra core speed, based on informal monitoring
- Total compute resources now **216 cores**, **704GB RAM**
- Development node (24 core, 256GB)
  - Currently being used for IPv6 testing
- Making more use of volumes (Cinder), for persistent storage
- Upgraded OpenStack to Icehouse
  - Component by component in-place upgrade now (mostly) achievable, with minimal downtime (cf. CERN blog post)
  - Upgrades included in CI testing by developers
  - API cleanups
  - Better IPv6 support

Imperial Cloud Status continued

- Still using Gluster for shared storage
  - Older hardware for Ceph problematic
  - Still plan to reinstate Ceph for long-term
- 2.5TB instance storage, 512GB volumes, 128GB images
- Database was becoming a bottleneck
  - Solved by removing old tokens (~1 million rows)
- Running mixture of CMS, ATLAS and LHCb jobs, plus local testing
- CMS using “Stealth Cloud” and Andrew Lahiff's glideinWMS

Imperial Cloud Plans

- Future plans:
  - Neutron networking
  - Ceilometer (accounting/monitoring)
    - » Will probably require dedicated database node
  - Docker (support limited at present)
  - Expansion...
- VMDIRAC (Simon Fayer & Daniela Bauer)
  - In testing at the moment
  - May need to use pre-release version owing to missing features in last public release (e.g. cloud-init)
  - Working on data management next
- Package building (cf. Fedora Infrastructure)

EGI Federated Cloud

- “a seamless grid of academic private clouds and virtualised resources, built around open standards and focusing on the requirements of the scientific community”
- Integration requirements (for OpenStack):
  - VOMS-enable Keystone installation and configuration
  - OCCI installation and configuration
  - Integration with accounting service APEL
  - Integration with VM Image Management infrastructure
  - Integration with information system
  - Registration of deployed services in GOCDB
- Both taken from EGI site

EGI Federated Cloud continued

- Necessitates change to Keystone, fundamental part of OpenStack (identity service), on which other services rely
- Instead of built-in server, use Apache with SSL enabled and mod_wsgi
  - Large performance impact
  - Makes debugging harder
- Serious bug in Keystone VOMS module
  - A lot of effort from Simon Fayer, including building a custom version of the VOMS library with extra debugging output
  - Eventually led to patch that fixed their handling of certificate chains
- Had to submit patches to support OpenStack Icehouse, for Keystone-VOMS and OCCI-OS
- Need a new BDII (didn't want to mess with existing ones)

EGI Federated Cloud continued

- Instructions for key parts (e.g. Keystone VOMS) very much skewed towards Debian
- Obscure configuration options that have no effect on normal operation can stymie FedCloud services e.g. Nova option for OCCI
- Accounting work ongoing
  - Current accounting code doesn't work on Icehouse and code is effectively abandoned (developer left in early 2013)
  - Intentions to use Ceilometer on OpenStack
    - » Some prototype code, not finished or ready for production
- Unclear if there are Puppet modules for these components (1 out of date Github repo)

EGI Federated Cloud continued

- Appears to have different motivations than GridPP
- 'Different pricing models under consideration' – Models for Sustainability (under development)
- Unclear yet how to marry VO and FedCloud usage patterns
  - Maybe better to handle in parallel cf. Oxford plan

Cloud Work around GridPP

- RAL
- QMUL
- Lancs
- Oxford

RAL

- Deploying OpenNebula
  - ~1,000 cores
  - ~1PB raw Ceph storage
- Initially private
- Funded by STFC, so primary use is for STFC/Scientific Computing Department projects (see Ian's talk)
- Will be closely coupled to the Tier 1, so bursting of unused resources will be possible
- Will also provide cloud endpoints for LHC VOs

Lancs

- Limited ongoing testing of VMWare
- Main focus on Vcycle and Vac (see Peter Love's talk)
- Aim to deploy more Vac factories and test Vcycle (see Andrew McNab's talk)

Oxford

- No change in existing cloud, running ATLAS jobs
- Installed Shoal on local Squid to avoid CVMFS thrashing at remote sites
- Now planning a new cloud to run in parallel, aiming to supplant current one in time
  - OpenStack Icehouse
  - Will add OCCI as step towards FedCloud integration
  - Hardware:
    - Dell R610 controller node, 16 cores, 24G
    - Network node running Neutron, 2 x 10G plus 4 x 1G NICs
      - Should mitigate network bottleneck caused by lack of multi-host support in Icehouse
    - Dell R510 storage node for Glance and block, Dell R510, 12 x 300G SAS
    - Mixture of new quad-core and old WNs as compute nodes

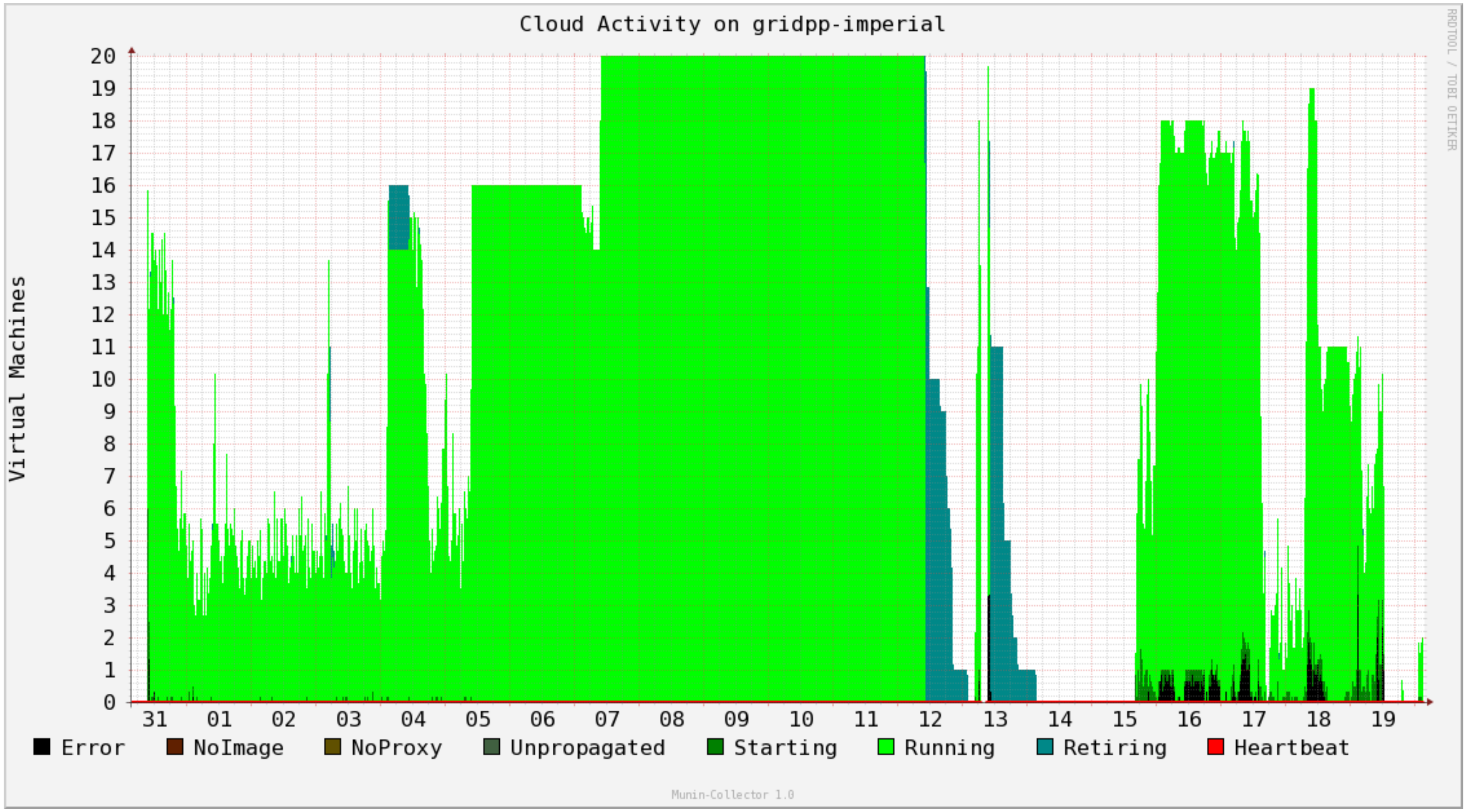
QMUL

- Running CloudStack
- Dell 1950 and PowerVault 3000 for storage
- Supermicro Twin Squared compute node
- Future plans:
  - EC2 does not work out of the box however there is a new project:
    - » https://github.com/BroganD1993/ec2stack
  - There is an option to try LXC containers and Xen alongside KVM, using different ZONEs.
  - CEPH / glusterfs file system instead of NFS
  - Based on usage, add more Twin Squared compute nodes
  - Add more storage

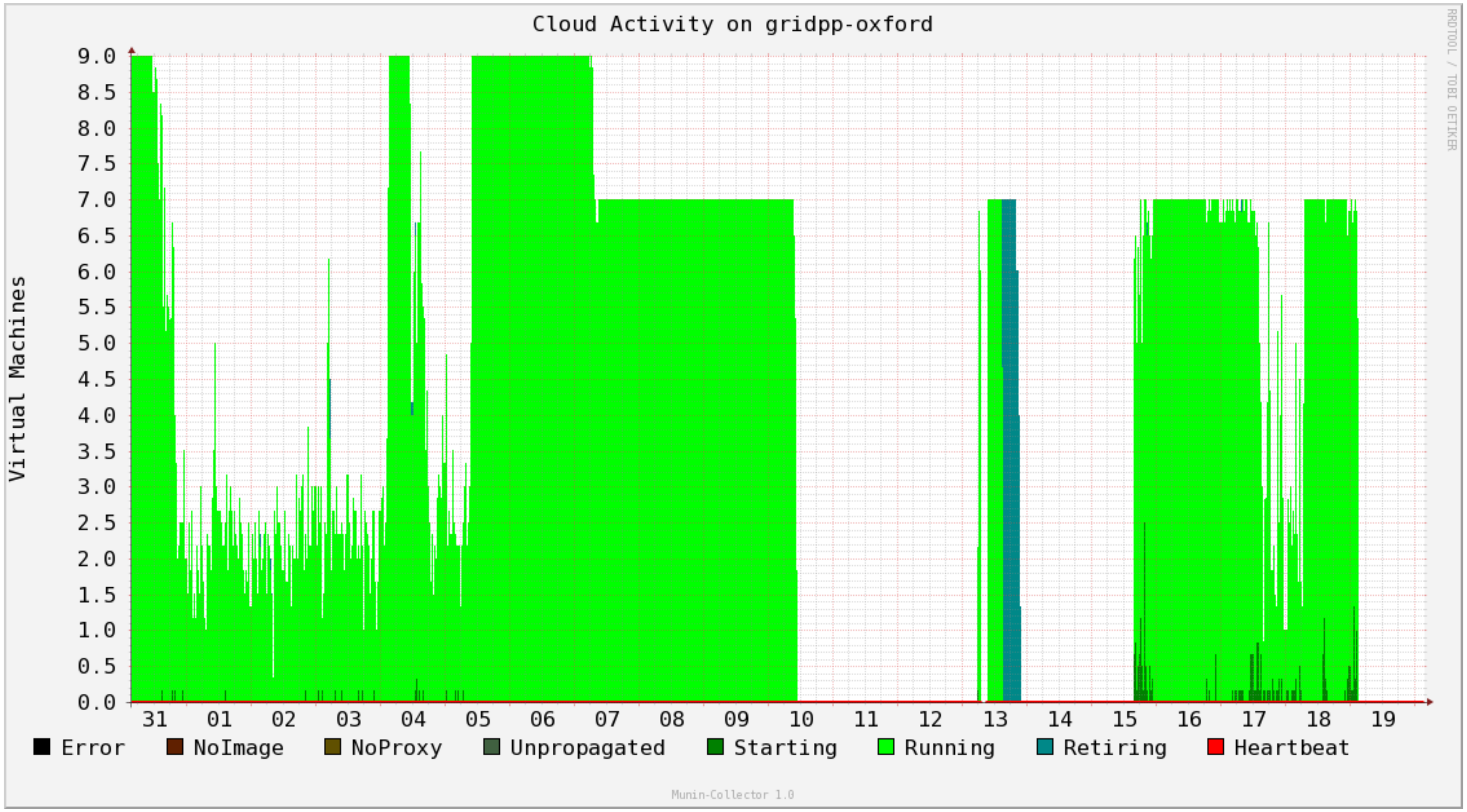
Cloud Usage

- Been useful for “real work” for quite a while now
- See Andrew Lahiff's talk for CMS information
- Using Simon Fayer and Daniela Bauer's “Stealth Cloud” at Imperial
  - By definition, CMS jobs running on this cloud aren't separate from normal grid jobs
- See Andrew McNab's talk for LHCb/Vcycle plots

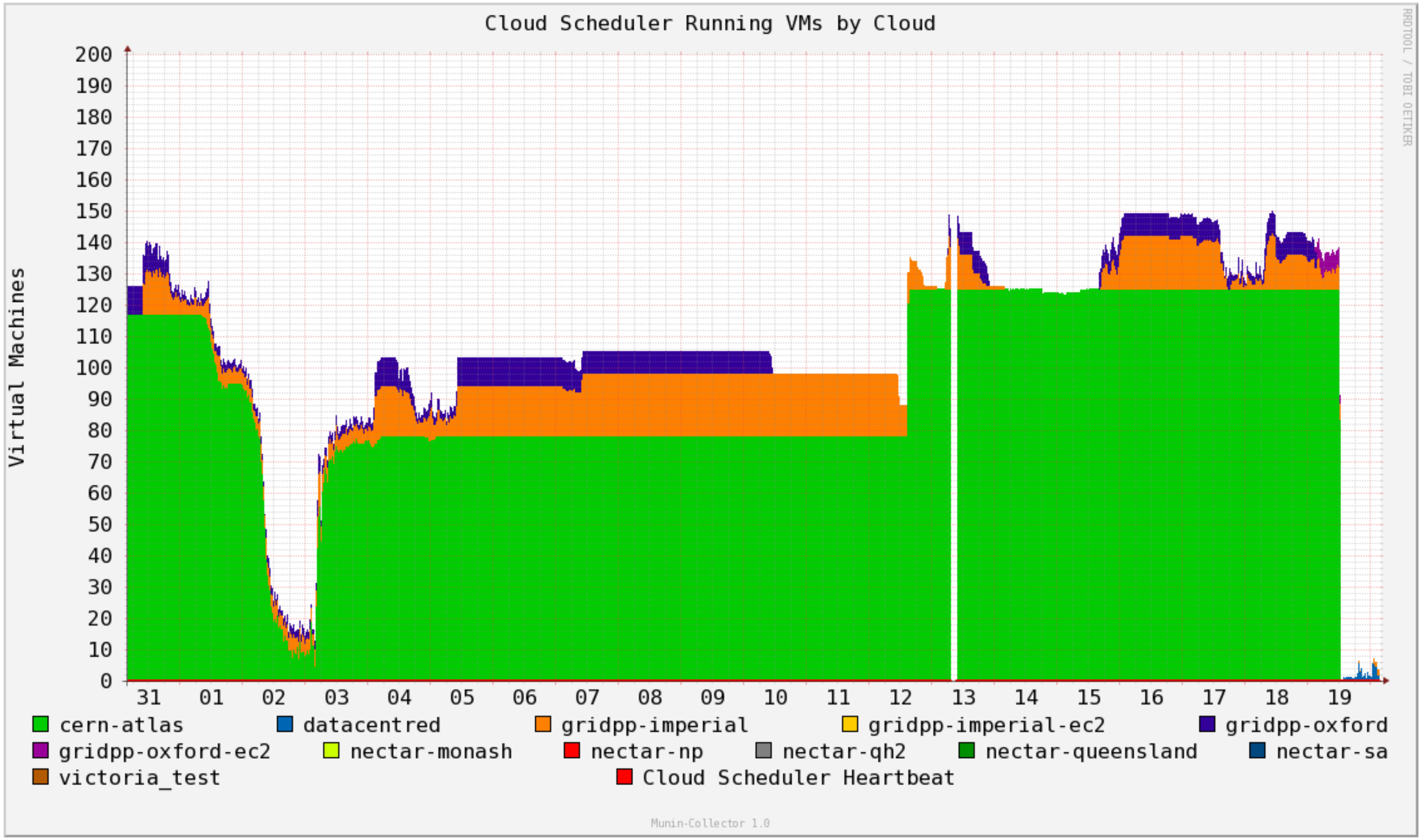
Usage – ATLAS at Imperial



Usage – ATLAS at Oxford



ATLAS Cloud Usage



Clouds in Production

- Bureaucracy needed as we move to production
- E.g. need to become a site registered in GOCDB
  - Procedures very much skewed towards grid sites (understandably)
  - In discussion with Jeremy about this
- There seems to be a conflation of certification as a site and certification in the federated cloud, though this may flow from the special case of 100%IT, which is the only test case that's completed so far
- Security questionnaire

Making better use of Clouds

- Wasted resources caused by flavours that are too large
  - Having a mix would be best from the sites' point of view
  - Eg. ATLAS uses 8 vCPUs, CMS uses 2 and Vcycle uses 1
  - How does this fit in with Grid multicore work?
- Dynamic quota handling to adapt to changing demand?
  - Can VO schedulers cope with this?
  - Current schedulers still don't cope with instances in the error state perfectly, though this has improved a lot
  - How would “demand” be quantified and communicated?
  - Sites monitoring turnover of instances/staleness?
    - Noticed that glideinWMS/CMS instances lingering a lot longer than ATLAS/Vcycle ones (may just be a default setting)
  - Cf. Andrew McNab's “target shares”

IPv6

- Appears to work straightforwardly on CERN OpenStack
- 'Officially' supported in nova-network
- Proposals to improve support in Neutron (which will be the main network service in the future)
- Testing at Imperial ongoing

Thanks

- Site informants:
  - Ian Collier
  - Kashif Mohammad
  - Robin Long/Peter Love
  - Dan Traynor