CoEPP

ARC Centre of Excellence for Particle Physics at the Terascale

# Australia Site Report

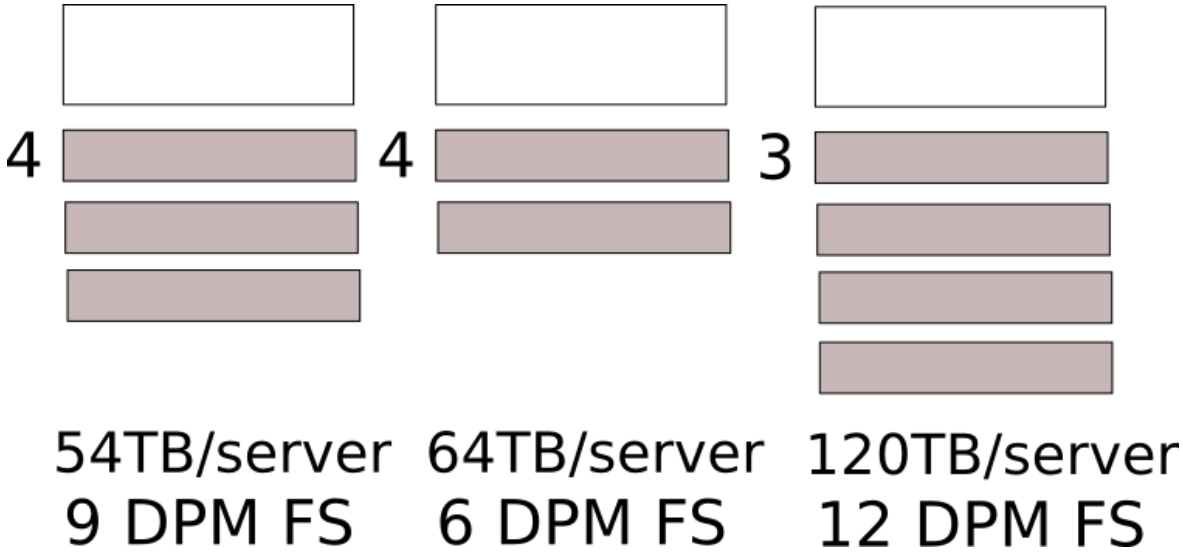
Sean Crosby

DPM Workshop – 13 December 2013

# Australia-ATLAS

- Supports ATLAS VO since inception
- Tier 1 site is TRIUMF (Canada Cloud)
- 10Gb WAN link through Seattle
- Aim to, and delivering, 2% of ATLAS compute
- 2 (soon to be 3) fulltime sysadmins
- 1000 cores
- 900TB storage
- Support 100 researchers (Academics, Postdocs, PhD and Masters)

- DPM 1.8.7/SL6
- 11 disk servers



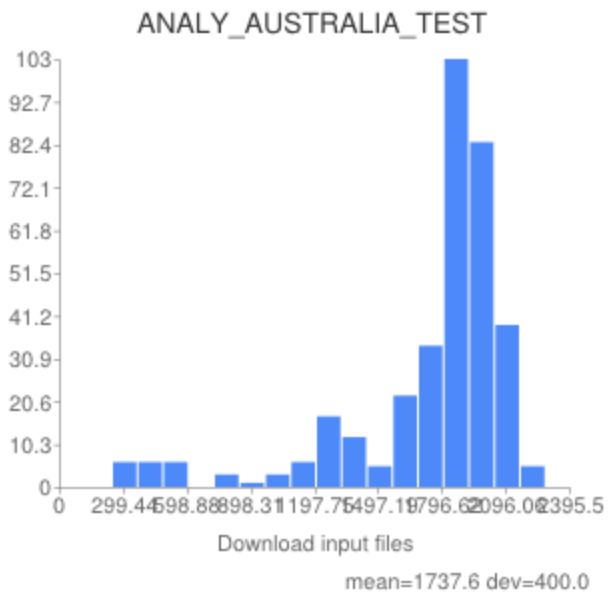

- 2x10GbE to switch (most compute nodes 1GbE)

- Standalone head node
- 1GbE
- SRMv2.2, xrootd, rfio, gsiftp, webdav, dpm, dpns
- Provides storage for physical cores (on our network), and Cloud cores (see later)
- Database moved to dedicated DB server – serves MySQL (MariaDB) and Postgres for all our other services
- Result – load on head node never exceeds 0.2!

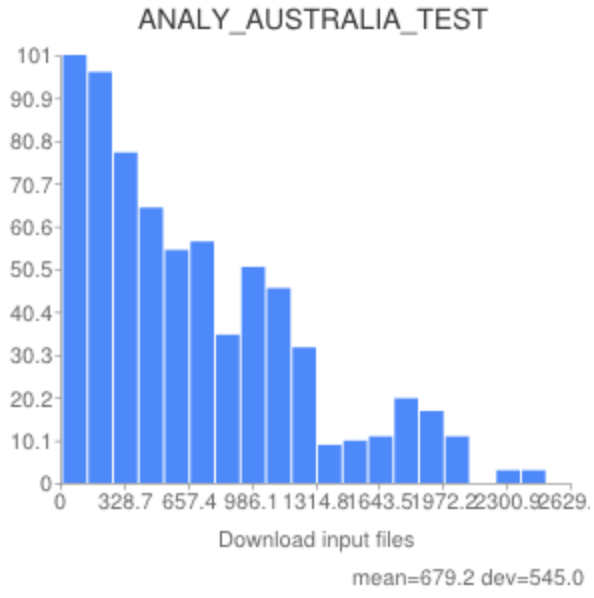
# Progress this year

- Moved head node to SL6/puppet control, and all storage nodes to SL6
- Drained 150TB from old nodes and decomissioned
- HammerCloud stress tested set up – before addition of new storage, on SL5, and on SL6
- Changed Cloud queues to use xrootd for stagein (was rfio)
- Removed firewall from TRIUMF link

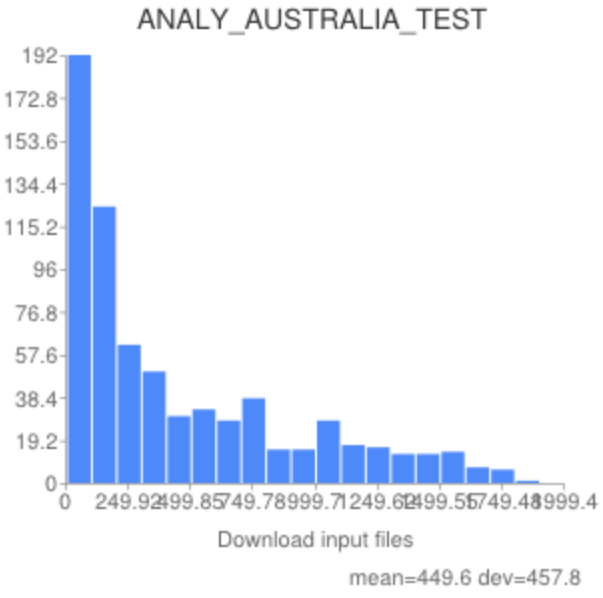
- Got great efficiency benefits, but also stage-in time




SL5 compute and old storage (some with 1GbE connections)




SL5 compute and new storage (all 10GbE)




SL6 compute and new storage (all 10GbE)

# HammerCloud tests

- Tests done using rfio for stagein
- Want to repeat the tests with SL6/xrootd stagein/xrootd direct-io
- Have to wait until next storage purchase, as HC tests use approx 15TB of datasets as input, and must be stored on LOCALGROUPDISK. No room!
- Once results are known, will switch physical cores to xrootd

# ATLAS and Webdav

- ATLAS requires sites to activate Webdav endpoint for Rucio namespace migration
- Once upgraded to 1.8.7 (and related lcgdm-dav), rename worked fine (see below for problems)
- Started on 6 November
- Completed 10 December

- When rename started, head node was swamped
  - dpm daemon log showed all new requests in DPM_QUEUED state
  - never processed
  - DB connections reached approx 600
  - netstat connections reached 30000 (didn't check number when dpm daemon wasn't processing requests though)
    - 98% of connections was in TIME_WAIT state to dpnsdaemon

- Changes made
  - NsPoolSize reduced to 12
  - net.ipv4.tcp_tw_reuse = 1
  - net.netfilter.nf_conntrack_tcp_timeout_established = 3600 (was 432000)
  - net.netfilter.nf_conntrack_tcp_timeout_time_wait = 15 (was 120)
  - net.ipv4.ip_local_port_range = 15001 61000
  - ATLAS reduced concurrent Webdav connections

- Result – no more DPM “hangs”

# Current issues

- After rename, all data in old, non-Rucio locations, is obviously dark
- How to find number of files and sizes?
- Tried gfal2 mount of SRM head node
  - Directories show up as 0 bytes in size
  - Will try gfal2 mount of xrootd/DAV head node next

```
data09_1beam
data09_2TeV
data09_900GeV
data09_cos
data10_1beam
data10_7TeV
data10_900GeV
data10_hi
data11_2p76TeV
data11_7TeV
data11_hi
data12_8TeV
mc08
mc09_7TeV
mc09_valid
mc10_14TeV
mc10_2TeV
mc10_7TeV
mc10_valid
mc11_14TeV
mc11_2TeV
mc11_7TeV
mc11_900GeV
mc11_valid
mc12_14TeV
mc12_7TeV
mc12_8TeV
mc12_valid
rucio
SAM
step09
user
user10
```

- Current set up is not balanced
  - drain of 8 old disk servers
  - 90% of data (140TB) went to the 3 new disk servers
  - Need dpm-addreplica API to support specifying specific disk node/FS

- Starting to support Belle2 VO
  - Started new GOC site Australia-T2
  - Will use storage provided for us in different cities in Aus

# Developments

- Storage will be attached to VMs (also provided to us) mostly via iSCSI on hypervisor (Openstack Cinder)
- We should get ~1PB
- Storage nodes near compute as well (also Cloud provided)
- Will also support ATLAS

- Current idea
  - Single SE
  - BELLEDATADISK and ATLASDATADISK token, distributed across all storage locations
  - MELBELLESCRATCH, ADLBELLESCRATCH, SYDBELLESCRATCH....., to ensure output is written local to compute node
  - Can DPM support this?
  - dCache

- Our Tier 3/batch cluster needs /home and /data storage
  - multiple locations
  - ~1000 cores (cloud provided)
  - Current idea
    - /home (~40TB) provided by CEPH-FS out of Melbourne
    - /data/mel, /data/adl, /data/syd writable using NFS on compute node, readable using local xrootd federation from any compute node in any location
    - Local WAN caches to reduce latency
    - Work in progress

# Thanks

- Wahid for xrootd ATLAS configs
- Sam and Wahid for troubleshooting and all round help
- David, Oliver, Fabrizio, Alejandro, Adrien for great chats at CERN this year
- Everyone on DPM mailing list

scrosby@unimelb.edu.au