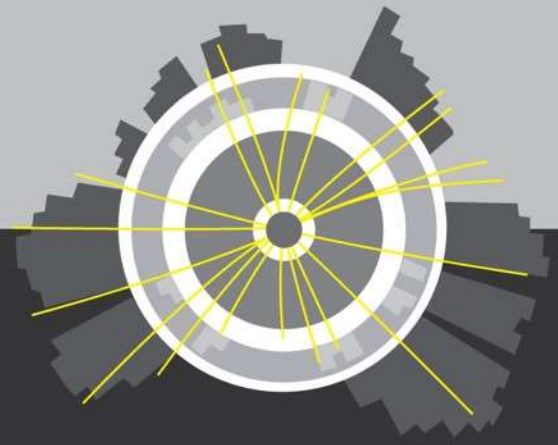


ARC Centre of Excellence for
Particle Physics at the Terascale

Federating Australian HEP Research Storage Using XRootD

Federated Storage Workshop
Sean Crosby
Australia-ATLAS
Melbourne, Australia

Acknowledgements

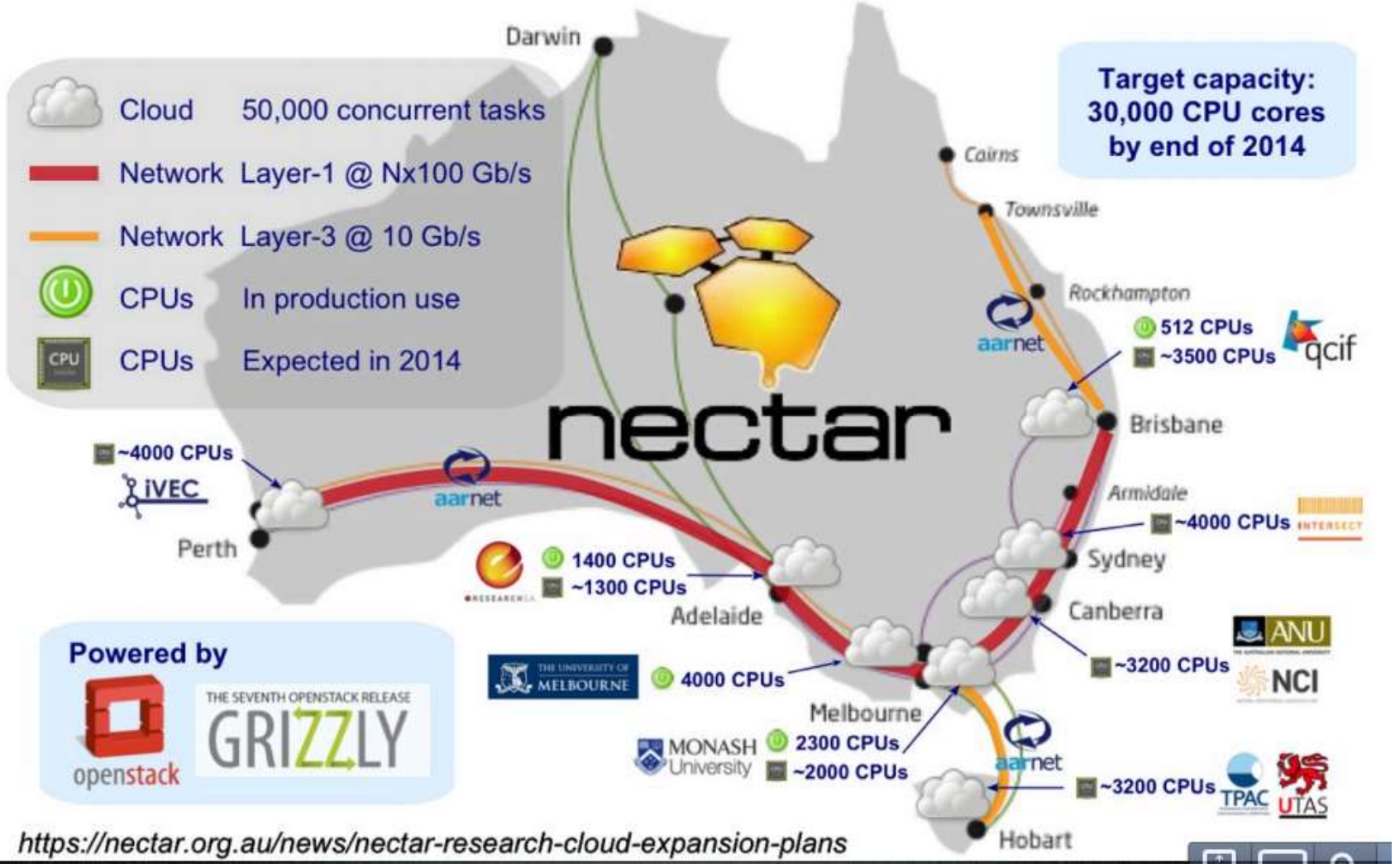
- Antonio Limosani
- Tristan Bloomfield
- Doug Benjamin
- Wei Yang

Research Computing Team

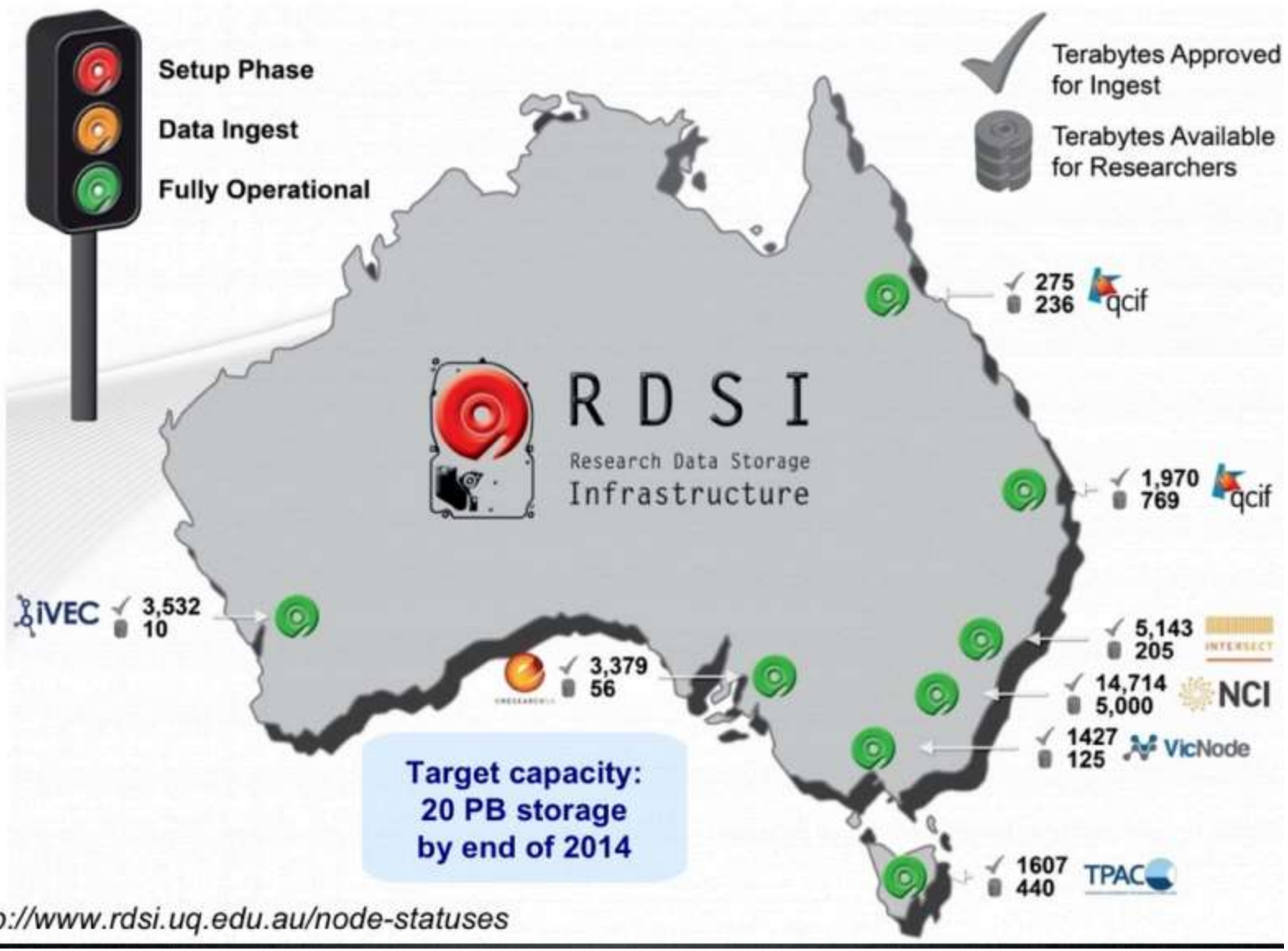
- Lucien Boland

- $25million over 7 years from Aus Government
  - Join the HEP groups from Uni Melb, Uni Syd, Uni Adelaide and Monash Uni together for first time
  - Also first time experimentalists and theorists were joined
  - Approx 80 FTE academics, postdocs, PhDs and Masters students
  - Research Computing group (2 members so far) to maintain Australia-ATLAS and the local systems
    - Purchase and deploy new pledge for ATLAS
    - Keep hardware in warranty for local systems

Other government money



Other government money



- Allocations are approved by a merit committee
- Factors include high importance, size of user community, how often dataset is accessed
- Clearly ATLAS compute and data fits all of these categories
  - We have been very successful in obtaining compute and storage
  - Have been allocated 700 cores and > 300TB so far (not all online yet)
  - 200cores used for Australia-NECTAR, 200 for Tier3, 200 for Belle2

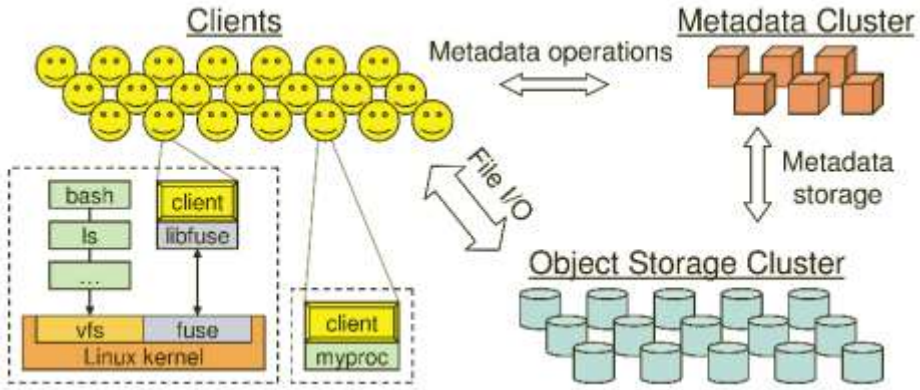
- We buy commodity hardware (Dell, IBM, HP) for compute and storage
  - Run compute until it dies
  - Decommission storage after 3 years
  - Rate of decommission approx 240TB/year
- How best to use Govt equip and our decommissioned hw?

- All network based
  - Mostly NFS in VM
  - Some Openstack Cinder (iSCSI terminated on hypervisor, block device in VM)
  - No dedicated storage network (with 1 exception)
  - Each site is different
    - Different SLA
    - Different speed and breakdown
    - Different functionality (backups, replication etc)
    - Individual LUN limits at some sites

Our plan for storage

- Need /home and /data
  - Separate them for backups
    - /home backed up, /data not (limited backup space)
  - /home for scripts, unrecoverable data
  - /data for DQ2 downloads, user-gen data
  - Approx 40TB for /home, infinite space for /data

- Use decommissioned hardware
  - RAID10 with 20% hotspares
  - Keep 30 drives for cold spares
  - CEPH (CEPHFS via FUSE)
  - Single location (Melbourne)
  - Mount on physical nodes, Cloud VMs
  - Working quite well so far
    - No major problems
    - Quite performant
    - Fault tolerant
    - Replica count = 2
  - To do
    - Get more users on
    - Install private network for replicas
    - Investigate SSD for journal



linux-mag.com

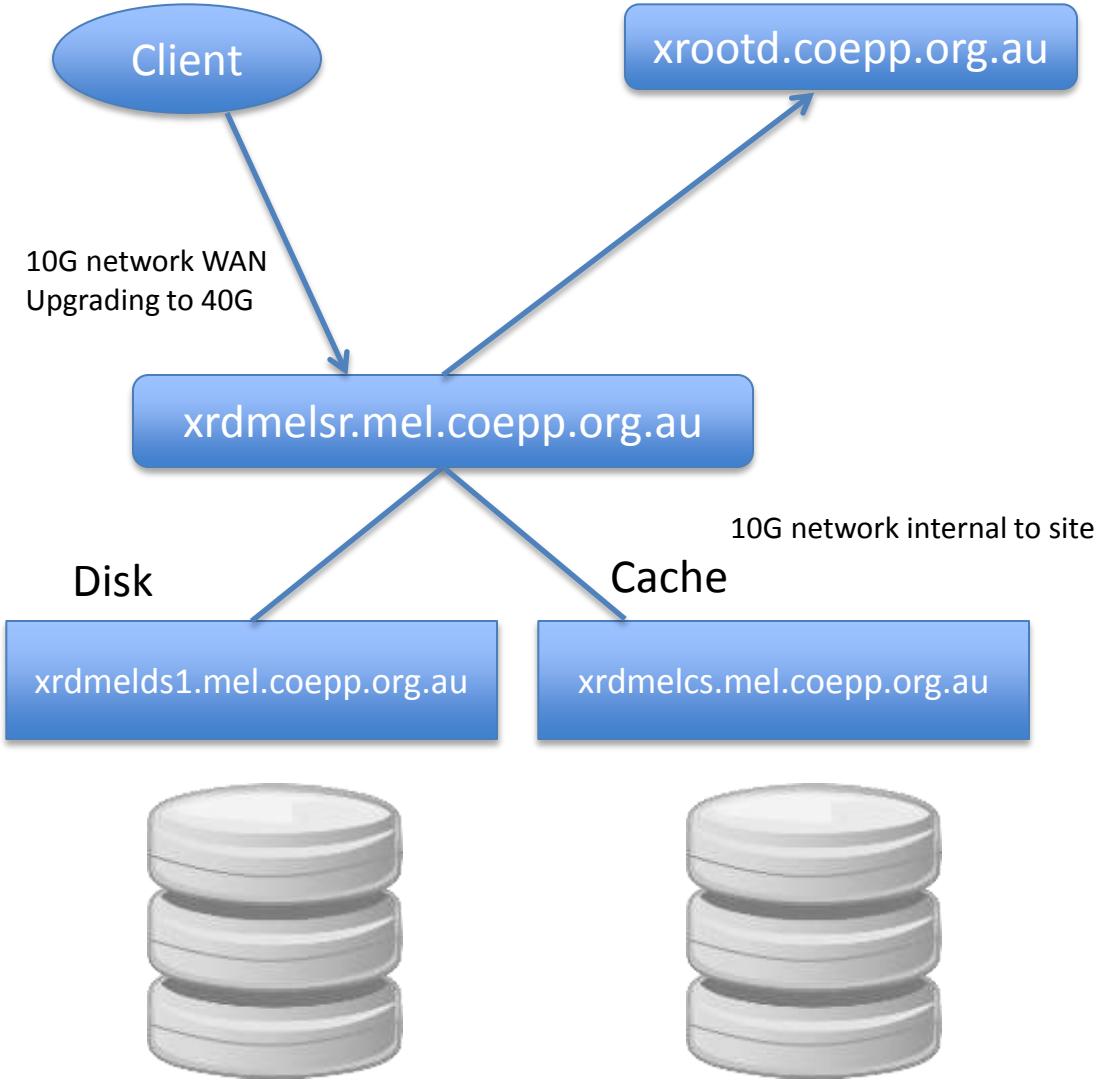
- Mostly experimentalists, but also non-neg theorists
  - Prefer POSIX-like FS
- Needs
  - Multiple sites
  - Pluggable
  - Performant
  - Fault tolerant
  - Not immutable
  - ROOT functionality a plus

- Lots of testing of distributed FS
  - Xtreemfs
  - dCache NFSv4.1
  - OrangeFS
  - FhGFS
- Most suffer from lack of reliability (Xtreemfs and OrangeFS especially), or lacks functionality (dCache – immutable – simply set up to test NFSv4.1 kernel speed)

xrootd

- Doug pointed us towards xrootd
  - Familiar with it from DPM
  - Initial configs from Doug and Wei
  - Initial idea for xrootd to be RO, writing done via NFS on WN
- Each site
  - Site Redirector (VM)
  - Disk server(s) (VM with NFS or Block storage)
  - Cache server (VM with NFS or Block)
- “Global” redir
  - VM
- Unix auth, xrootd user in LDAP, with appropriate group permissions (atlas, belle)

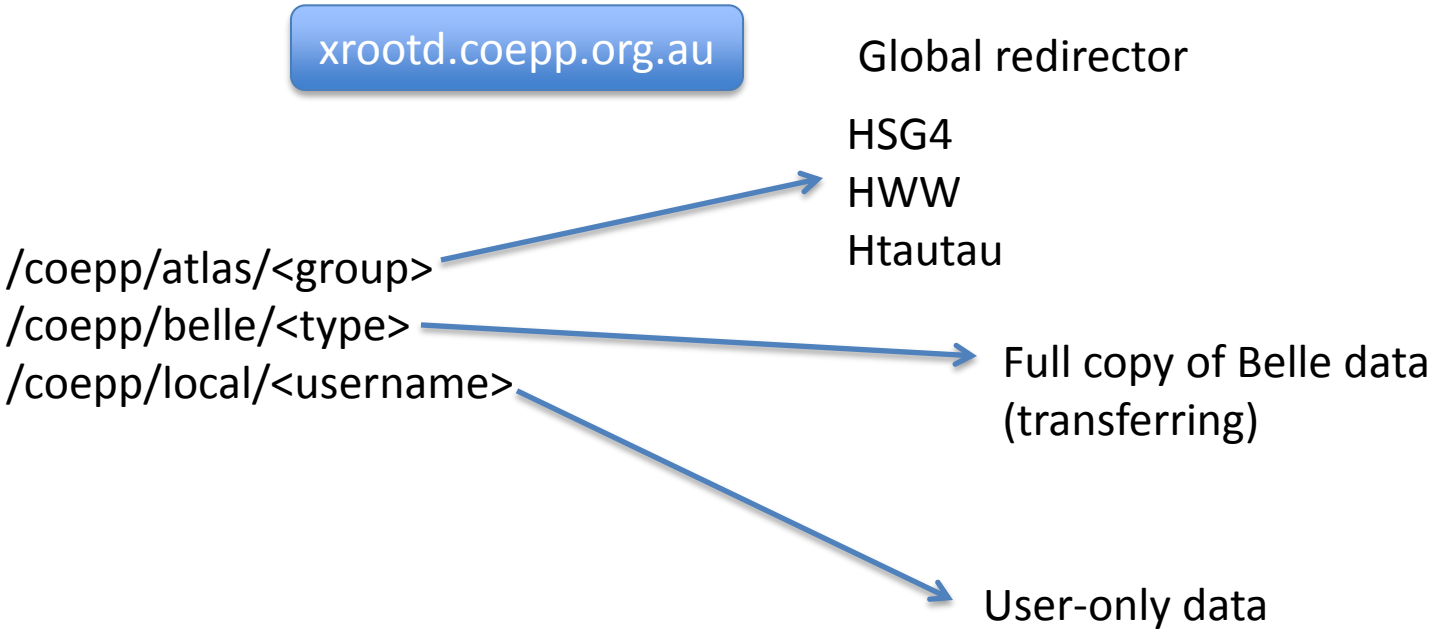
xrootd



Global redirector

Same basic setup replicated to other sites using same puppet configs

Namespace


xrootd.coepp.org.au
Global redirector
HSG4
HWW
Htautau
/coepp/atlas/<group>
/coepp/belle/<type>
/coepp/local/<username>
Full copy of Belle data
(transferring)
User-only data

Initial results

- ROOT analysis job
- Input : 7 GB dataset containing 90K LHC ttbar events stored in TTree
- Output : histograms
- Cache turned off (site level and TTree)
- Results have yet to be replicated (they don't make much sense to me)

	Melb Disk	Syd Disk	Adl Disk
Melb CPU	00:13:35	02:49:12	04:08:23
Syd CPU	03:08:18	00:28:18	06:19:18
Adl CPU	03:55:09	06:06:22	00:38:08

- Same job as before

	Syd Disk	Adl Disk
Syd CPU	00:28:18	00:32:37
Adl CPU	00:42:35	00:38:07

- Clearly cache works, but not as we like or expect
  - Xrootd cache server responds that it has the file, even though it doesn't
  - Stage-in script (provided on Twikis) had bugs (fixed)
  - Copies the file in, then gives it to the client
  - Copy problems result in inaccessible file
  - Given the network between sites is great, is that best?

- Turn off site caches
- Repeat with 100MB TTreeCache

	Mel Disk	Syd Disk	Adl Disk
Mel CPU	00:08:43	00:29:31	00:17:34
Syd CPU	00:23:34	00:08:51	00:22:30
Adl CPU	00:20:09	00:29:49	00:09:02

- TTreeCache is much more important
  - Will keep the cache servers, but will reevaluate

Problems

- FUSE
  - Xrd FUSE mount extremely slow
    - ls takes O(mins) to finish
    - Need cns?
      - Cns confused by NFS writes
- Enable xrootd writes
  - Melb DS had data already
  - Not in new directory structure
    - Tried to force it by config change on that DS
      - Oss.localroot: disk space reporting wrong
      - All.export: xrdcp would segfault across federation
- Unresponsive SR or DS caused slowdowns for everyone
- Syncing DS directories a problem
  - Mel now has 3 DS (due to LUN size limits)
  - Xrd mkdir only mkdir on individual DS

Further Work

- Next step to implement cns and FUSE mount
- Been investigating pyxrootd
  - Get around most problems with theorists?
- Education
  - Tier3 and Tier2 level – our DPM has been xrootd enabled for ever
  - Stop the double download
- Migration of existing data

WebDAV

- Fed WebDAV (Fabrizio UGR) is very exciting for us
  - Davix in ROOT big advantage
  - Dynamic federation
  - Browse dir structure using browser
  - Standards (protocol and servers)
- Will install apache/mod_dav/ugr in cohabitation with xrootd for near future

Thank You

scrosby@unimelb.edu.au