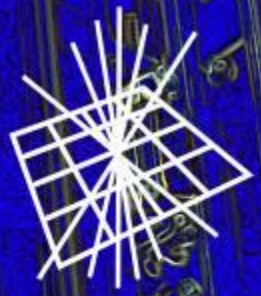

GridPP

UK Computing for Particle Physics

CEPH at the Tier 1

Brain Davies

On behalf of

James Adams, Shaun de Witt & Rob Appleyard

Science & Technology Facilities Council

Introduction

- What's Changed
- CEPH – the natural choice?
- What Tier 1 is doing
- Costs
- Risks

What's Changed?

- Previous analysis showed no imperative to move to a new 'disk-only' system
  - Little gain for a lot of pain
- But life and politics move on…
  - CERN seriously considering running Ceph under CASTOR for tape
  - Some issues previously identified in Ceph are, or will soon be, addressed
    - Erasure encoding, stability of CephFS
  - Fundamental design differences between CASTOR and XrootD are causing issues…
    - XRootD is very aggressive for access. Does not time out and can cause performance hits if clients are not responsive
  - CERN also starting to contribute to Ceph code base

Why Ceph?

- CASTOR want to drop current file system support
  - If Ceph works as well as planned
- Gets us out of Particle Physics specific software
  - Except CERN are contributing to the code base
- Improved resilience
  - Currently loss of 3 disks on server will (probably) mean loss of all files
  - Under Ceph, 3 disk loss will lose less (not quantified)
    - Assuming suitable erasure encoding/duplication
      - 2 erasure encoded disks per 16 physical
- Improved support
  - Ceph also planned for Tier 1 and SCD cloud storage
  - More cross-team knowledge

Plans and Times

- Currently developing quattor component
- Plan is to deploy ‘small’ test instance for all VOs
  - 1Pb nominal capacity, less overhead
  - Initially using CephFS and dual copy
  - Move to erasure encoding as soon as possible
  - **NO SRM**
  - Anticipate deployment late April/early May
  - Integration of XRootD RADOS plugin as soon as available
- After some months of VO testing (Oct. 2014?)
  - Start migrating data from CASTOR to CEPH
  - Need to work with the VO to minimise pain
    - Fewest possible files migrated
  - Start moving capacity from CASTOR to Ceph

ALL DATES SUBJECT TO CHANGE

Costs

- Ceph with dual copy too expensive long term
  - Need erasure encoding
- Could deploy with current set-up
  - 1 copy, RAID6, 1 hot spare
  - … but not recommended
    - Lose advantage of disk loss
- With erasure encoding…
  - Single erasure encoded copy (w/o hotspare, 1 erasure disk per 17 data disks) is cheaper than current setup
    - But less resilient
  - Dual erasure encoded copy (w/o hotspare, 2 erasure disks per 16 data disks) is about the same price
    - And better resilience

Proposed Deployment

- Ideally ‘single’ instance with quotas…
  - Single meaning 1 disk instance and 1 tape instance (still under CASTOR)
  - Using Ceph pools
  - Simpler to manage rather than 4 instances currently set-up
  - Easier to shift space around according to demand
- Problem may be ALICE security model
  - May force us to run with 2 instances
  - Work with ALICE to see if this can be mitigated

Risks

- Lots of them...
  - This is just a sample

Risk	Likelihood	Impact	Mitigation
Erasure encoding not ready	Low	High	None
CephFS not performant	Medium	Medium	Switch of http access
CephFS not stable	Low	High	Switch of http access
XRootD/RADOS plugin not ready	Medium	High	Use POSIX CephFS
Difficulty in migrating data	High	High	Minimise data to be migrated
Difficult to administer	Medium	Medium	Use testing time to learn about the system
Ceph moves to support model	Low	Low	Buy support from Inktank (or other vendor)