Contribution ID: **66** Type: **Oral Presentation**

Experiences of hard disk management in a erasure coded 10 petabyte-scale Ceph cluster

Tuesday, 20 March 2018 15:00 (30 minutes)

RAL has developed a disk storage solution based on Inktank and Red Hat's 'Ceph' object storage platform to support the UK's WLCG Tier One centre. This solution, known as 'Echo' (Erasure-Coded High-throughput Object store), is now providing ~40% of the disk storage available to the LHC experiments and wider WLCG at the UK Tier 1.

Data is stored in the cluster using erasure coding to reduce the space overhead while keeping the data safety and availability high in the face of disk failures. An erasure coding profile of 8 data shards plus 3 erasure coding shards is used to provide data safety levels comparable to a system with 4-fold replication at a fraction of the cost, keeping the space overheads down to 37.5%.

In large-scale Ceph clusters, deployment of storage nodes without RAID controllers is preferred, mainly for performance reasons. This comes at a cost however, as any and all disk errors become visible to Ceph and will cause inconsistencies to be noticed that must be dealt with, generating administrative workload for the service's operators. Factors from disk age and disk placement in the physical machine to disk utilisation and load contribute to create and operational environment that requires constant attention and administrative intervention.

In this paper, we present the history of disk-related problems that have affected Echo. Echo consists of approximately 2200 disks with a raw capacity approaching 13 PiB. A certain quantity of disk problems are to be expected when running this quantity of hard disks. However, we have also experienced a problem where undetected bad sectors on disks have a destructive interactaction with a Ceph bug that can trigger cascading failures during routine data movement operations. We present methods the Ceph team have utilised to cope with this workload, as well as how the disk management tooling provided by the Ceph project has evolved.

Primary authors: Mr VASILAKAKOS, George (STFC); Mr BYRNE, Tom (STFC)

Co-author: Mr APPLEYARD, Rob (STFC)

Presenter: Mr APPLEYARD, Rob (STFC)

Session Classification: Data Management & Big Data Session

Track Classification: Big Data & Data Management