Echo update

Alastair Dewhurst, George Vasilakakos, Ian Johnson, Bruno Canning, James Adams, Alison Packer

Introduction

- Echo status
  - Erasure Coding
  - CRUSH maps / Failure Domains
- Multiple Slash Collapse
- GridFTP work to be covered by Ian next

echo

Echo status

- We start deploying our Echo cluster with the new hardware two weeks ago.
  - Various differences with SL7 slowing us down.
- 3 Mons and 30 storage nodes currently up.
  - Another 30 storage nodes to do.
  - No benchmarking done – Will report next month.
- Spent some time looking at cluster configuration
  - CRUSH Map.
  - Erasure Coding.

Erasure Coding

- Other people (CERN? + Yahoo) are using 8 + 3 Erasure Coding in production.
  - We have enough raw capacity for 8 + 3.
  - Can revisit higher EC later but currently focusing on plugins.
- Which EC code to use?
  - Jerasure is Ceph default – probably because it works with any architecture.
  - CERN use ISA?, which is better for writes but only works on Intel CPUs[1,2,3].
- Thoughts from the community?

[1] http://ceph.com/planet/ceph-jerasure-and-isa-plugins-benchmarks/
[2] http://arxiv.org/pdf/1504.07038.pdf
[3] https://indico.cern.ch/event/524549/contributions/2149939/subcontributions/195700/attachments/1289767/1920265/cephhepday-dan.pdf

CRUSH Maps

- Current plan is to ensure all OSDs in a placement group are on different storage nodes.
- Is it worth building larger failure domains into the CRUSH map?
- With EC 8+3 we would need to have a minimum of 4 failure domains
  - Each failure domain would have at most 3 OSD in it.
  - Would the cluster still actually function with the loss of ¼ of the storage nodes?
- Opinion:
  - The CRUSH map should be designed around potential permanent data loss.
  - Large failure domains (e.g. network switches) should be configured redundantly rather than rely on Ceph.

echo

Multiple Slash Collapse

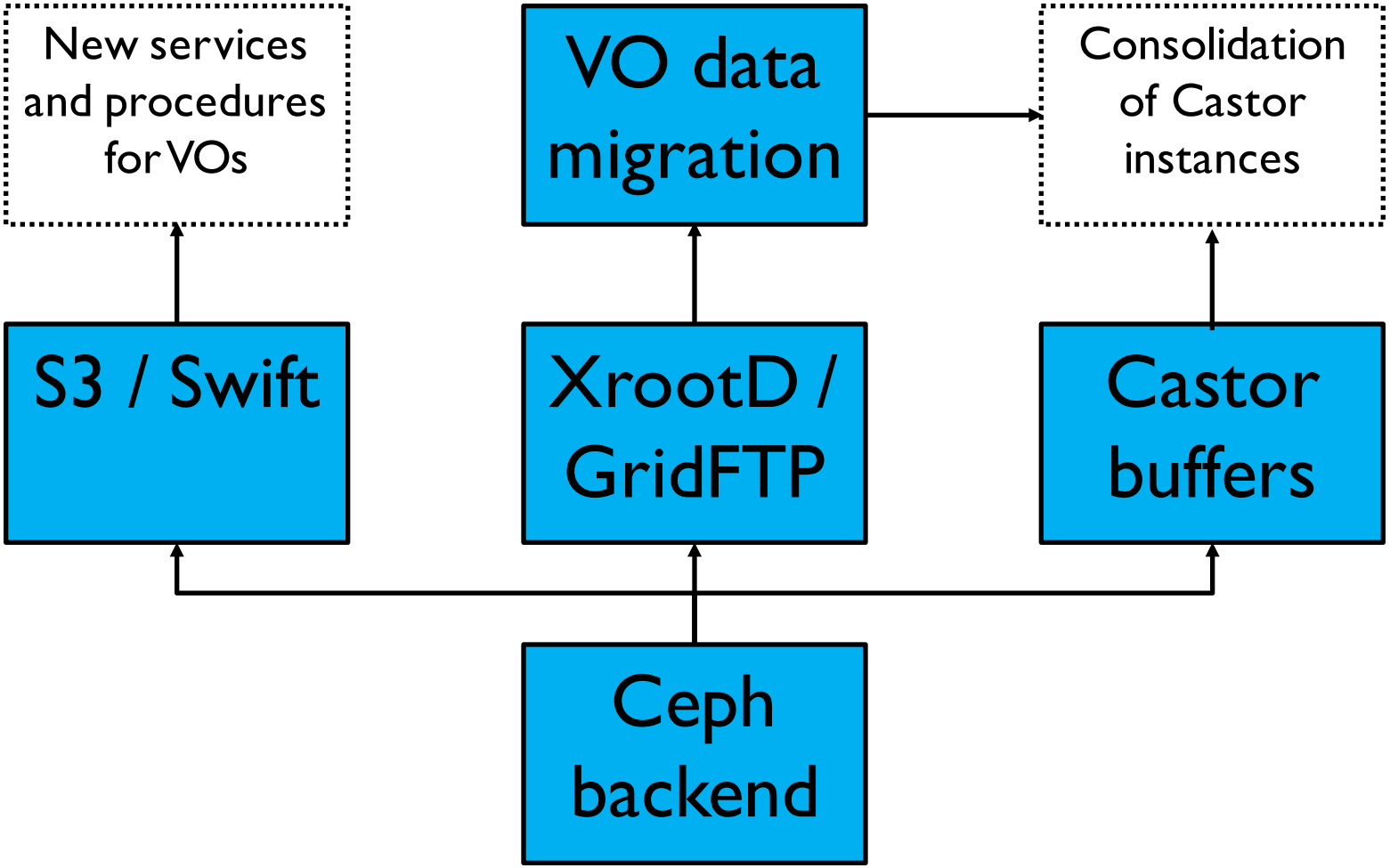
- Both the XrootD and GridFTP plugins will collapse multiple slashes.
- Not very object store like behaviour
  - We intend to keep it because it is likely to break VO workflows.

```
-bash-4.1$ xrdcp root://lcgvo05.gridpp.rl.ac.uk//atlas/rucio/user/ivukotic:user.ivukotic.xrootd.ral-lcg2-1M /tmp/deleteme
[1024kB/1024kB][100%][==================================================][102.4kB/s]


-bash-4.1$ xrdcp root://lcgvo05.gridpp.rl.ac.uk//atlas/rucio/user////////////ivukotic:user.ivukotic.xrootd.ral-lcg2-1M /tmp/deleteme
[1024kB/1024kB][100%][==================================================][341.3kB/s]
```

Backup

The Echo Project



Erasure Coding

- Typical Castor disk server with 36 drives, 30 are storing data.
  - 2 disk for OS, RAID60 (i.e. 2 x 15+2) = 83% of raw storage is usable.
- EC breaks data into 'k' chunks and creates 'm' additional parity chunks.
  - Can lose any 'm' chunks without losing data.
- For Ceph we want m = 3 (at least).
  - Allows us to take advantage of self healing (reducing effort required to maintain).
- To keep overhead down, need k to be as large as possible without affecting performance.

m \ k	8	10	12	14	16
2	80	83	86	88	89
3	73	77	80	83	84
4	67	71	75	78	80

percentage of raw storage usable

Yahoo

Facebook

Tier 1 goal

Hardware

- 3 × monitor nodes: Dell R420, RAM: 64GiB, CPU: 2 × Intel Xeon E5-2430v2, 6 core, 2.50GHz.
- 3 × gateway nodes: Dell R430, RAM: 128GiB, CPU: 2 × Intel Xeon E5-2650v3, 10 core, 2.30GHz.
- 63 × storage nodes: XMA (Supermicro X10DRi), RAM: 128GiB, CPU: as gateways, OS Disk: 1 × 233GiB SSD, Data Disks: 36 × 5.46TiB HDD (WD6001F9YZ) via a SAS HBA.
- Total Raw Storage = 12.1PiB, 13.6PB.

echo