Contribution ID: **25** Type: **Oral Presentation**

8 Years and 100PB of Ceph: How We Do It, Lessons Learned, and Future Plans

UKRI-STFC's Scientific Computing Department operates Echo, a very large (~100 PiB) data storage cluster located at the Rutherford Appleton Laboratory and implemented using Ceph storage technology. Echo is one of the world's largest publicly-advertised Ceph clusters. Echo's primary use case is to provide a high-throughput data storage endpoint for the WLCG's high-throughput computing operations - it serves data to RAL's Condor batch farm, to remote batch farms at other UK institutions, and to other WLCG sites.

After a brief overview of Ceph and the RAL Tier 1, we will give an overview of the cluster's development, design goals, and the cluster's current state. The specific highlights include:

Typical hardware configurations of the storage, gateway, and monitor nodes

The specific approach we use for WLCG file storage - user files are striped to 64MB objects, which in turn are stored using an 8+3 erasure coding scheme.

Gateway configuration: high bulk throughput is achieved through tight integration between the local batch farm and the storage cluster.

We then move on to an overview of the challenges that managing a Ceph cluster at this scale poses. Specific issues with very large clusters include node rebuild times, monitor node load, internal networking capacity, data migration times, and the duration of major interventions. We address each of these in turn and detail our approaches to each.

Finally, we will outline our recent changes and future plans for the cluster and general Ceph provision at RAL. With 260 storage nodes, Echo has surpassed the default host-level failure domains, necessitating a more sophisticated architecture - we refer to this project as 'High-level failure domains'. We discuss the constraints on this architecture imposed by the combination of data centre infrastructure (power, networking, physical rack layout), and Ceph's own architectural requirements.

Primary authors: Mr MCCOMB, Aidan (STFC-UKRI); Mr TOM, Byrne (STFC-UKRI); Mr APPLEYARD, Rob (STFC-UKRI); Mr ABUAJAMIEH, Maksim (STFC-UKRI)

Presenter: Mr APPLEYARD, Rob (STFC-UKRI)

Track Classification: Track 7: Network, Security, Infrastructure & Operations