Science and Technology Facilities Council

Ceph at RAL

HEPiX Spring Workshop 2021

15-19 March 2021
Morgan Robinson,
STFC UK Research and Innovation

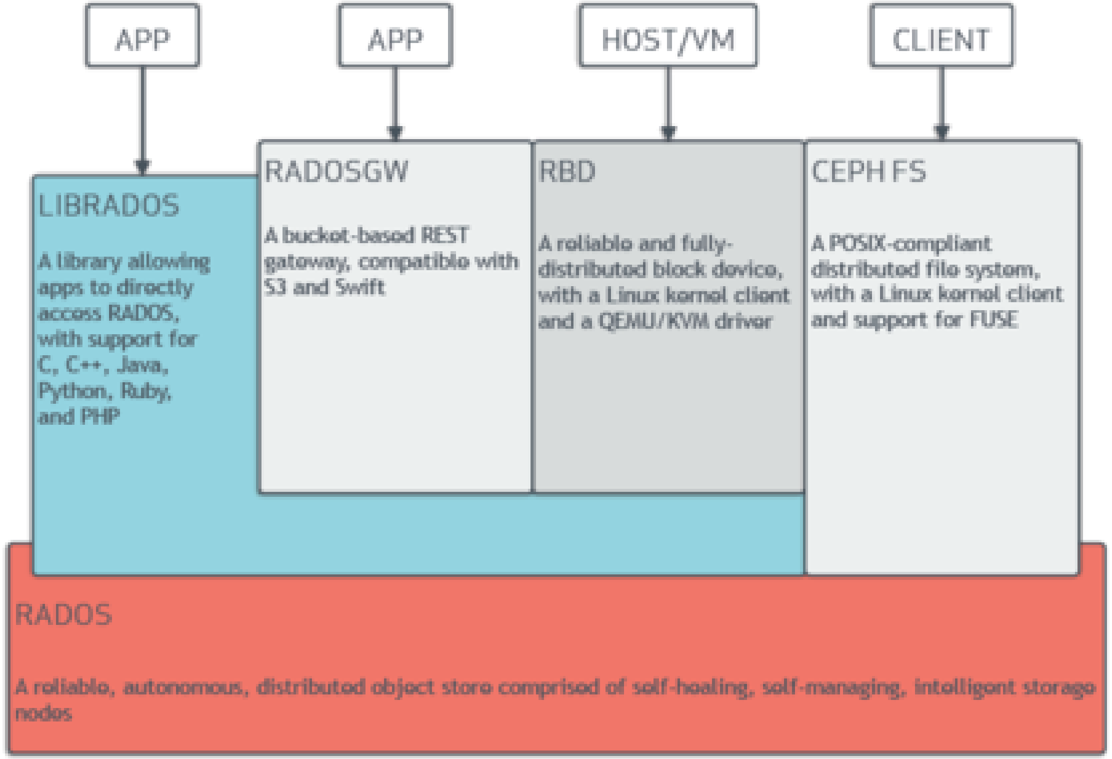
Computing at RAL

At RAL we have multiple users and clients that require storage solutions for their projects:

- STFC Facilities
  - ISIS - Our Muon source installation
  - CLF – Central Laser Facility
- STFC Cloud Team
- GridPP – Tier 1 Storage for the LHC experiments

What is Ceph?

- To help provide for these users we use Ceph on our clusters.
- Ceph is an open-source software storage platform, which provides interfaces for:
  - Object storage
  - Block storage
  - file-level storage
- Ceph is highly scalable.



A library allowing apps to directly access RADOS, with support for C, C++, Java, Python, Ruby, and PHP

A bucket-based REST gateway, compatible with S3 and Swift

A reliable and fully-distributed block device, with a Linux kernel client and a QEMU/KVM driver

A POSIX-compliant distributed file system, with a Linux kernel client and support for FUSE

A reliable, autonomous, distributed object store comprised of self-healing, self-managing, intelligent storage nodes

Our Clusters

Within RAL we operate 3 separate storage clusters that all provide different services for users. All of which are currently running Ceph Nautilus (v14)

1 Echo

Our main object storage cluster, Echo has a high-capacity object store that we provide to external users including the LHC experiments

2 Deneb

Our CephFS cluster that is used by STFC facilities such as ISIS and CLF

3 Sirius

Block storage device cluster that is used by virtual machines for our SCD Cloud

Image © STFC Alan Ford

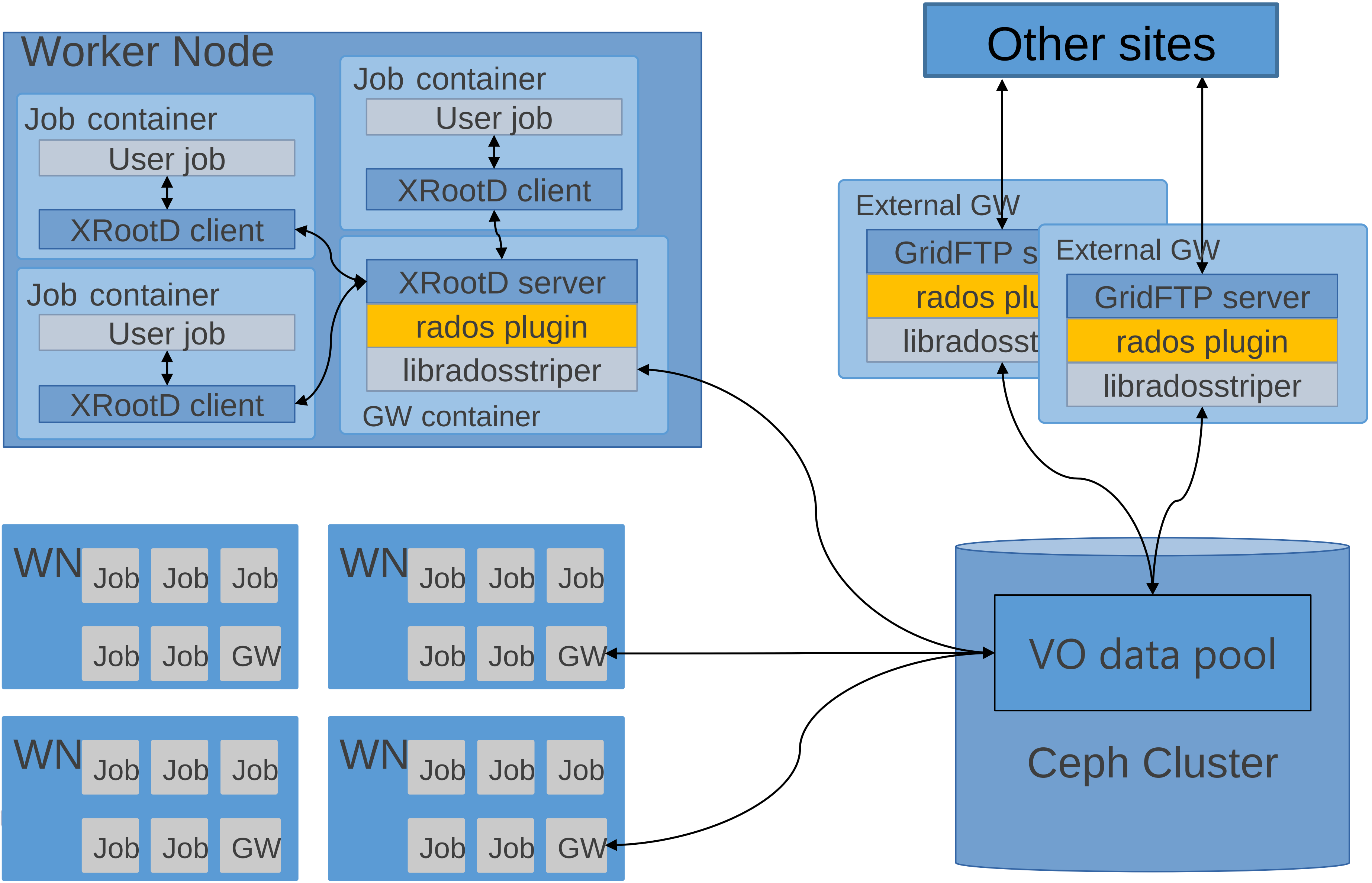
What is Echo?

Erasure-**C**oded **H**igh-throughput **O**bject store (ECHO) is the largest Ceph cluster at RAL with the main purpose of providing large scale, high throughput storage to support the LHC experiments as well as other groups. Access to the cluster is managed through using GridFTP, XrootD and S3.

ECHO uses 8/3 erasure coding:

- Each object (stripe) is split up into shards
- 8 data shards, 3 parity shards
- We can survive losing 3 whole storage nodes

Echo Architecture

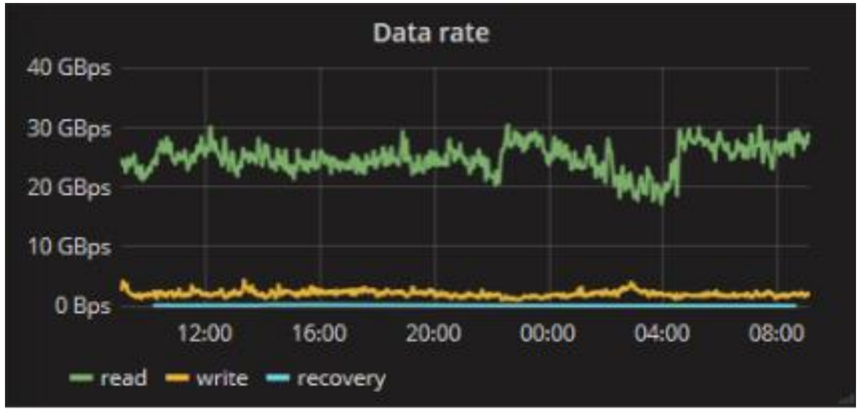


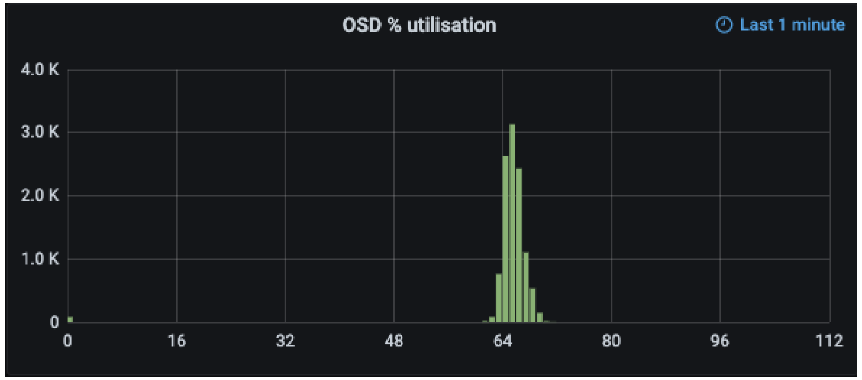
Science and Technology Facilities Council

Operational Details

Echo is running:

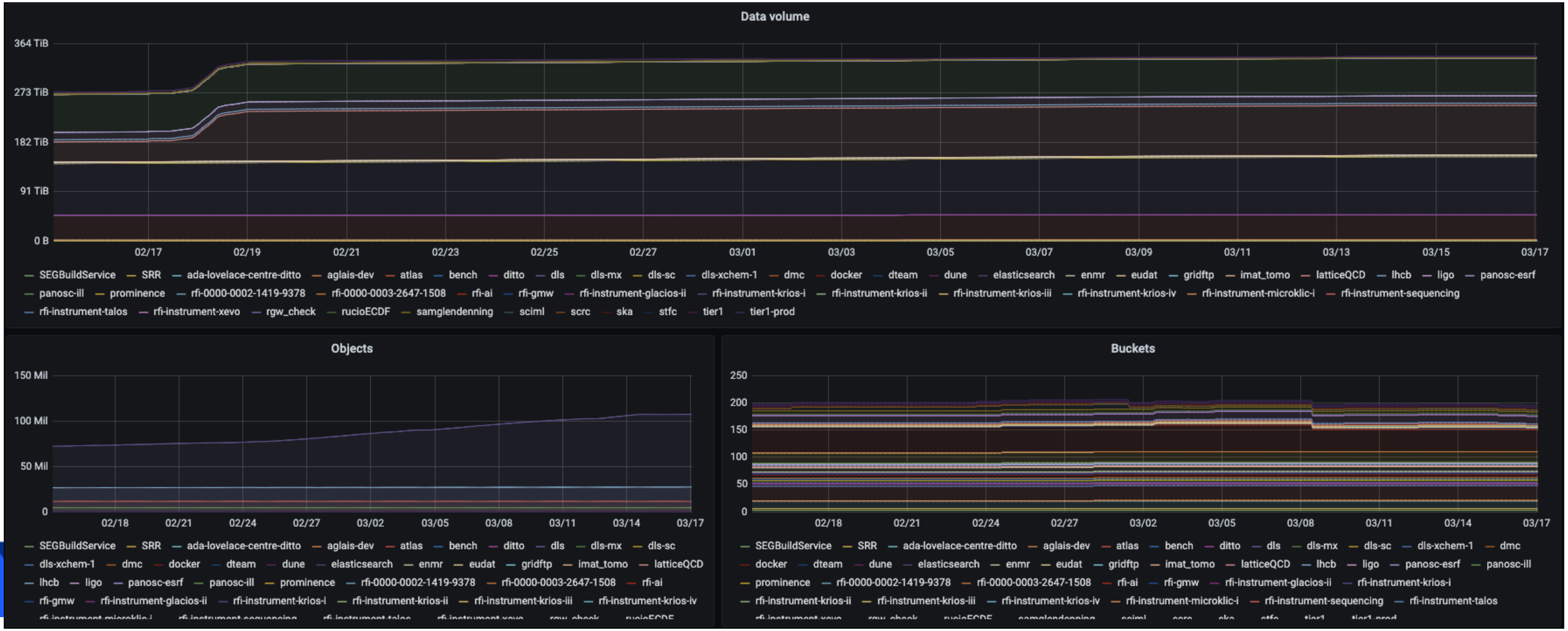
- 229 Storage Nodes w/ 5900 HDDs (Object Storage Daemon)
  - Providing 54PB raw space and will be adding hardware to take this up to 70PB.
  - Largest publicly known EC cluster
  - Average read rates 20 - 30GB/s
- 5 Monitors
- 11 External Gateways
  - 7 general purpose
  - 2 Alice - to accommodate different authorisation
  - 2 CMS AAA - to control throughput





Echo S3

- Usage of S3 endpoint is steadily growing (non-LHC community).
  - 340TB
  - 151 Million Objects

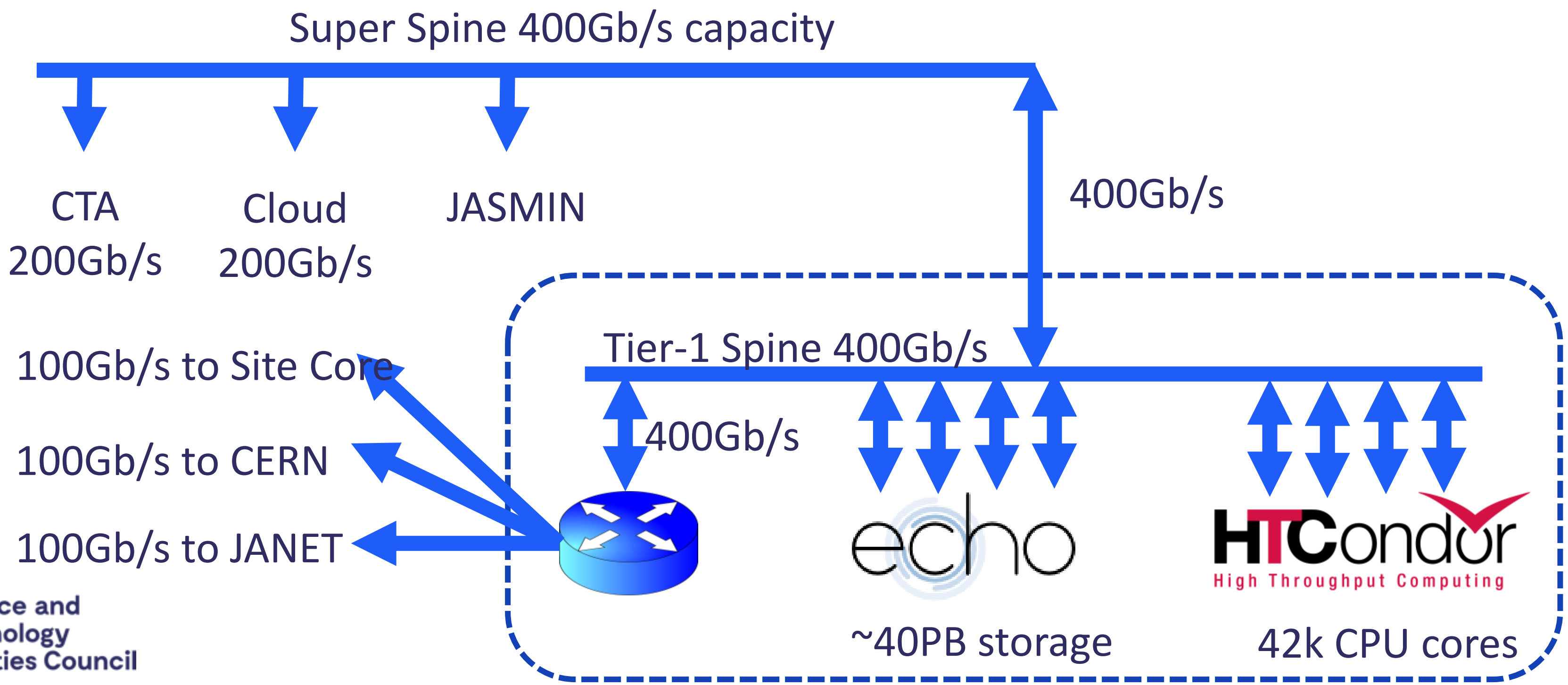

Data volume

Objects

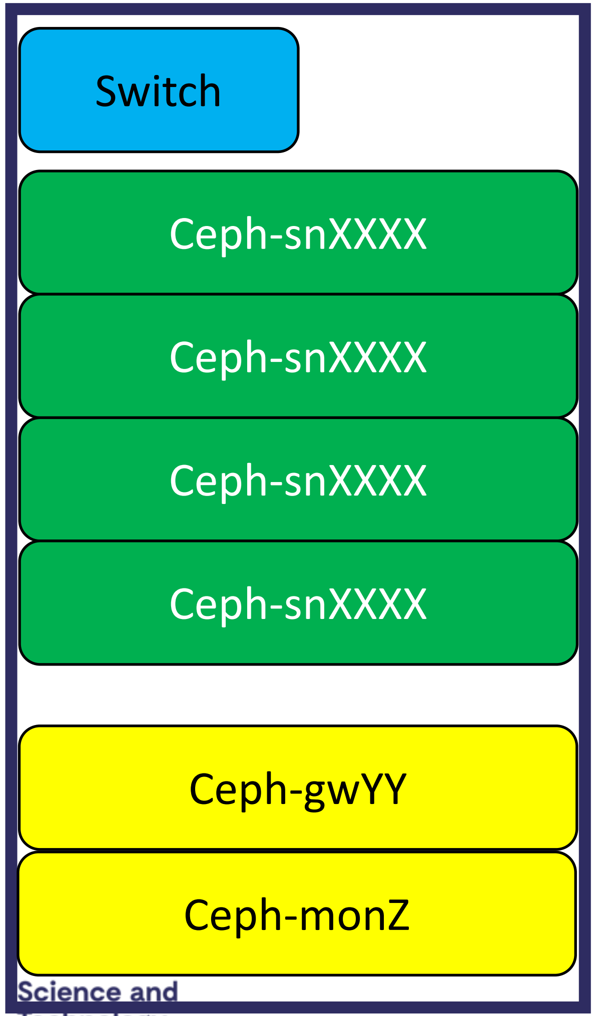
Buckets

Plans for Echo

- Tier-1 is upgrading network to a Spine/ Leaf topology.
- Will be able to provide non-blocking network for Ceph.



Echo rack design



- Each Rack will have
  - Top of Rack (aka Leaf) switch.
  - 10+ Storage nodes.
  - 1-2 Gateways / Monitors
- Each server will have a single 25Gb/s connection to the ToR switch.
- The ToR switch will have 4 x 100Gb/s connections to the Tier-1 Spine.
- Echo will eventually have ~20 racks like this.

Science and Technology Facilities Council

Deneb

Deneb provides CephFS file storage used by SCD facilities including on site projects and experiments such as ISIS and CLF Octopus.

CephFS allows us to create mountable filesystems within Ceph with each pool having it’s own file system. With this we can account for space usage, rates and enforce hard quotas

In the future we plan to use Deneb to support Openstack Manila that will allow us to provide file shares as a service for the STFC cloud

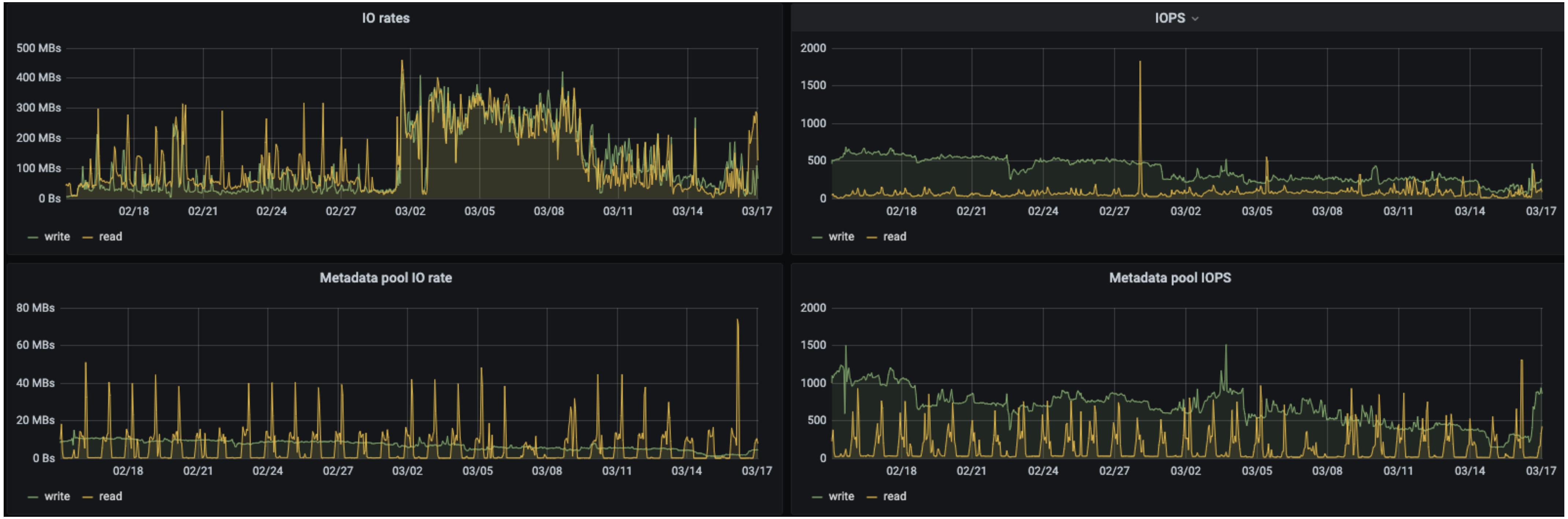
Deneb

About the Cluster:

- 42 Storage Nodes – 624 HDDs OSDs
  - Each storage node has 1-2 additional SSDs that contain the metadata store for the OSDs
  - By moving the metadata operations off the spinning disks this helps with IOPS
- 4 Metadata Servers
  - 2 Active + 2 Standby
- 4.0 PiB Raw space
  - Also using Erasure Coding 8+3 like Echo

UKRI Science and Technology Facilities Council

Deneb Performance



IO rates

IOPS

Metadata pool IO rate

Metadata pool IOPS

Sirius

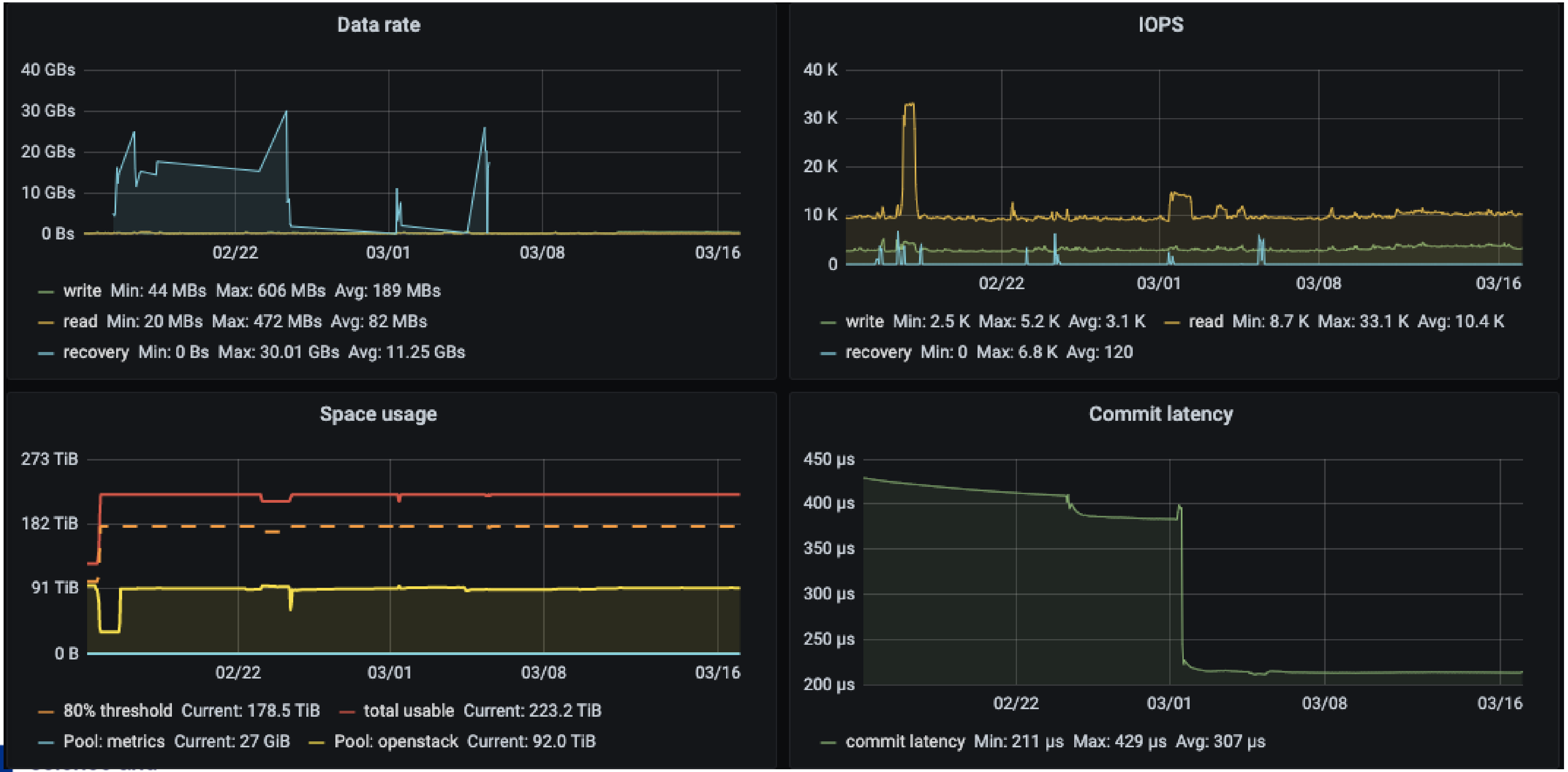
Sirius provides block storage for running cloud virtual machines that provide services for the SCD including some monitoring systems for the Ceph clusters.

Originally when it was created, Sirius was comprised entirely of HDDs but in 2020, this was changed as a demand for faster reads was needed for running VMs. All of the storage nodes in the cluster since then have been replaced with Solid State drives

About Sirius

- 23 Storage Nodes
  - 184 SSD OSDs
  - 670 TiB Raw
- Sirius uses 3 x replication.
  - Maximizing IOPs compared to EC.
- Recently added12 additional storage nodes
  - Additional 85TiB usable space.
  - Alleviate some issues within the cluster where we were tight on space if a machine had problems.

Sirius Performance



Data rate

— write Min: 44 MBs Max: 606 MBs Avg: 189 MBs

— read Min: 20 MBs Max: 472 MBs Avg: 82 MBs

— recovery Min: 0 Bs Max: 30.01 GBs Avg: 11.25 GBs

IOPS

— write Min: 2.5 K Max: 5.2 K Avg: 3.1 K — read Min: 8.7 K Max: 33.1 K Avg: 10.4 K

— recovery Min: 0 Max: 6.8 K Avg: 120

Space usage

— 80% threshold Current: 178.5 TiB — total usable Current: 223.2 TiB

— Pool: metrics Current: 27 GiB — Pool: openstack Current: 92.0 TiB

Commit latency

— commit latency Min: 211 µs Max: 429 µs Avg: 307 µs

Technology
Facilities Council

Lockdown Operations

Working from Home has introduced a few more challenges for operations:

- Restrictions to Machine Room Access
- Communication changes
- Possible delays in interaction
- Issues with monitoring systems

Ceph Clusters are capable of “self healing” which is extremely beneficial should we have any issues. If a storage node crashes, the cluster will adjust for the machine down and relocate files reducing the chance of data loss should additional machines fail

What next?

- When Echo was originally started the design was to have a large amount of nodes with 24 - 36 disks.
  - This is what we were used to with Castor.
  - Allowed us to spread data across storage nodes more easily (when we didn’t have many).
- With the scale Echo has reached we want to try a higher density machine to the cluster named, which we’ve named Tau:
  - 960Tb raw HDD storage
    - 80 Hard Disk Drives
    - 3x normally what’s on a storage node
  - 38Tb SSD storage

This will be a research node as well, to see what method of disk partioning would work best for the future

Potential Benefits

With Tau, this is the first time we’ve introduced SSDs onto Echo which gives us some options to experiment what to do with it:

- Different Tiers of storage
- S3 metadata operations
- OSD metadata stores

By opting to use higher density machines in the future this will provide us with the benefits of lowering future costs

Conclusions

Ceph’s adaptability has allowed us to meet the varied needs of our users. Instead of using just one general cluster to meet requirements, we’ve created multiple clusters over time each with a specific focus.

Ceph’s ability of self-healing has been beneficial providing ease of mind while working throughout lockdown and we’re quite happy to have had no major problems ☺

Science and
Technology
Facilities Council

Thank you