Federated Access from DOE Labs to Distributed Storage in the EIC Era of Computing

Michael D. Poat, Jérôme Lauret, Tejas Rao

May 11th, 2023
Outline

► Introduction
► Evolution / MinIO-Gateway
► Ceph S3
► IO Performance Tests
► Perspective / Federated ID
► Conclusion
Introduction

❖ The Electron-Ion Collider, a new facility for nuclear physics research to be located at Brookhaven Lab (BNL) but a cross-collaboration between BNL & Jlab

❖ While the computing model for the EIC is not finalized, we envision to have the storage resources accessible to a wide range of collaborators. This calls for
  ➢ A Federated storage solution
  ➢ A Federated ID access to the storage

❖ CEPH provides flexible ways to Federate storage (multi-location, pools with replication methods etc …) as well as the S3 protocol integrating Federated ID.

❖ As part of a “Program Development” funding, we established an S3 demonstrator / proof of principle.

❖ Our Deployments:
  ➢ Initial implementation used Lustre with MinIO-Gateway (Minio GW) for S3 access - test S3 use in EIC
  ➢ Second stage: A Ceph Object Storage with dedicated RadosGW Endpoints / S3
Our initial deployment was on a single host running the MinIO GW service with Lustre mounted underneath (dtn01).

Lustre setup (3PB):
- 3 Hosts : 48 Core, 392GB RAM, 4 x 25 GbE
- 100 x 14 TB per host -> 10 x 10-drive RAID6 OSTs

MinIO GW (v. RELEASE.2022-08-11) provides an S3 interface to GPFS / NFS / Lustre storage as a backend - quick and easy to set up.

It served its purpose, supporting the EIC detector design phase and was a stunning success. Accessible from anywhere, broadly accessible on the grid, BNL/S3 was the only read/write storage accessible.
A more robust setup followed using 4 hosts (28 Core, 132 GB, 4x25 GbE)

- 2 hosts running HAProxy/KeepAlived for failover and balancing
- 2 hosts for running MinIO GW/Lustre mount
- Resilience, fail-over, IO increase

Setup works well for a single site, but does not support zoning, Federated ID, and cannot scale across datacenters. Our goal is to provide a Federated access to Federated storage (storage could be added from anywhere)

Additionally, in 2020 MinIO GW over standard FS was announced to be deprecated and moving toward pure Object Storage as a focus. There was no path to continue with Lustre. Any evolution would need to support Object Store.
Ceph S3

- Ceph is a reliable and scalable storage system based on RADOS (Reliable Autonomic Distributed Object Store) - had experience with Ceph [M. Poat, J. Lauret – “Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques: A Case Study with CephFS”, (2016). (CHEP 2016)]

- It provides high availability and data protection, with features such as erasure coding and replication

- The Ceph Object Gateway is the interface built on top of librados providing the RESTful gateway between the storage clusters and the Amazon S3 API.

- OpenID Connect Provider in RGW – Federated ID Access is possible

- **Initial deployment:**
  - 3 - RadosGW: 28 Core, 256 GB RAM, 4x25 GbE
  - 4 - OSD Hosts: 48 Core, 96 GB RAM, 4x25 GbE, 8x16 TB OSD each

- ~450 TB RAW w/ Erasure Coding 4+2 pools (300 TB usable)

- Deployment is easily scalable, can add disks to current nodes or scale horizontally (add more storage nodes), infrastructure in place for scale out
A multi-site Ceph cluster can be configured as Multi-Realm, Multi-Zonegroup, or Multi-Zone.

Replication / mirroring / backups possible as some of the actions you can perform between sites:

- **Mirroring:** Mirrors two Zones across sites. Can write Objects back to either Zone, but all metadata must be written to the master - full sample at 2 sites.

- **Backup To/From Zones:** Directional Zone backup across sites (as read-only on secondary location).

- **Sync To/From Buckets:** Directional Bucket backup or Bucket Mirroring.

- **Sync To/From Buckets with Filter:** Sync Objects that match name regex to/from Zone/Bucket (*.daq, *.root, …)
Federated ID Access / STS for Ceph

- S3 authenticates with an ACCESS_KEY & SECRET_KEY not a secure method for distributed access

- **STS:** Secure Token Service is a web service that returns a Token & a temporary set of credentials for authenticating federated users. [STS in Ceph Object Storage - Pritha Srivastava (RH)]
  - Token contains the AuthN/AuthZ to the RadosGW (Roles - Who can assume a role & Role Permissions)
  - **AssumeRoleWithWebIdentity:** Used for any external application that wants to access S3 resource
    - Does not require owning any permanent credentials in S3
    - Users authenticate w/ external OpenID Connect/OAuth 2.0 compliant IDP

- Federated Access for our Ceph
  - We are currently in progress of implementing this - (was not fully functional in time for the conference [hardware delivery delays] but detailed configuration will be provided)
  - The OpenID connect provider in RGW should enable us to enable Federated Access. This is the key to unify cross-collaboration
Ceph S3 vs MinIO GW Single client Write tests.

► S3bench IO perf tool used for testing
  ○ 1 MB - 100 MB chunk size
  ○ 1 - 100 Threads

► Despite drive and size discrepancy, Ceph outperforms our MinIO GW setup among all chunk sizes.
► MinIO GW: Lustre is not using striping
► Single Ceph client can saturate it’s outgoing 25 GbE link with intensive writing.

► Ceph S3 vs MinIO GW - Multi (3) - client write tests.
► Ceph peak perf: 4.3 GB/s - Peak 100 MB chunks @ 25T
► MinIO GW peak perf: 2.1 GB/s - Peak 100 MB chunks @ 25T
MinIO GW vs Lustre backend - 2,000 clients

MinIO GW S3

► Submitted 2000 batch jobs using S3bench - 10,000 - 10 MB chunks per host (1 Thread per Endpoint (x2))
► MinIO GW: Peak write performance ~1.8 GB/s (on par with isolated tests)

2 MinIO GW Endpoints

Lustre Backend

► IOZONE IO performance tool used
► 2000 jobs Writing 10,000 - 10 MB chunks per host
► Peak Aggregate IO Throughput: 4.09 GB/s

Peak MinIO GW write performance at ~50% of the underlying of our Lustre storage. Our Lustre is tuned for read performance over writes.
Ceph Object Storage (S3) - 2,000 clients

- Seagate 16TB Exos X16 - **Manufacture Spec:** Max. Sustained Transfer Rate: 261MB/s
- Theoretical Aggregate Raw Speed: **8.1GB/s** (31 disks, 1 failed w/o replacement)
- Ceph Erasure Coding 4+2: **Theoretical performance:** 66% of Raw speed
- Submitted 2000 jobs: S3bench Writing - 10K - 10MB chunks per host to Ceph S3
  (1 Thread per Endpoint \(x3\))
- Peak Ceph performance: ~5.5 GB/s (68% of Raw Speed w/ EC 4+2)
- Performance in line with what we expect

- Ceph balances the IO among all disks within the cluster
- With a failed disk on one host, we can decipher the IO down to the disk level
- ~181 MB/s max throughput per disk with EC 4+2
- We can use this as a baseline to scale towards any aggregate IO requirements
A single Realm Multi-Site Ceph Object Storage provides a global object namespace and ensures unique object ID’s across the cluster.

While our Object Storage is focused on Ceph, a full MinIO Object Storage implementation or other object storage with S3 could be tested as options. Ceph provides a familiar technology and as solid baseline.

By vertically scaling: maximizing disks in current setup from 31 -> 48 disks: ~8.7GB/s Max Throughput.

By horizontally scaling: Double the number of hosts w/ 12 drives per: ~17.4GB/s Max Throughput.

The IO requirements of the EIC are not yet finalized but …

The scalability of the Throughput is however predictable.
Conclusion

Takeaways

► S3 accessible storage provided to the EIC production workflow - MinIO GW required minimal setup to have a globally accessible storage space (but lacked key features)
► Federated S3 Storage could provide the EIC with a globally accessible storage space that can easily be scaled locally (and has the ability to provide a framework where storage can be added from anywhere).
► Ceph Object Storage provides all the features to deploy a multi-site S3 storage.

Future/Ongoing work

► Implement and test the OpenID Connect module within Ceph
► Deploy an additional standalone cluster in a different location to test Multi-Site and the synchronization features
► Provide the skeleton/framework of a Multi-Site Ceph cluster with Federated ID access.
Thanks!
Backup

- 6 Month Grafana Chart: Lustre Read/Write