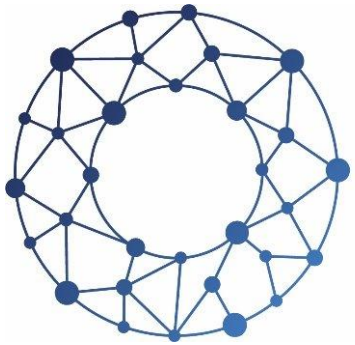


PSDI

PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Data transfer, synchronization and sharing solution for PSDI

Amali Pawula Hewage, Vasily Bunakov, Alexander Belozerov, Paul Wright, Stuart Meacham, Tom Underwood

19-21 March 2025

CS3 2025 Conference, LMU Munich

About Us

STFC (Science and Technology Facilities Council)

STFC is a part of UKRI (UK Research and Innovation). We operate large-scale facilities: ISIS neutron and muon source, Central Laser Facility, also co-own Diamond Light Source (synchrotron)

STFC do research beyond facilities:
in Particle Physics Department, Space Department, Scientific Computing Department

More about STFC Scientific Computing:
https://www.sc.stfc.ac.uk/

PSDI (Physical Sciences Data Infrastructure)

STFC Scientific Computing co-deliver PSDI with a few other partners in the UK

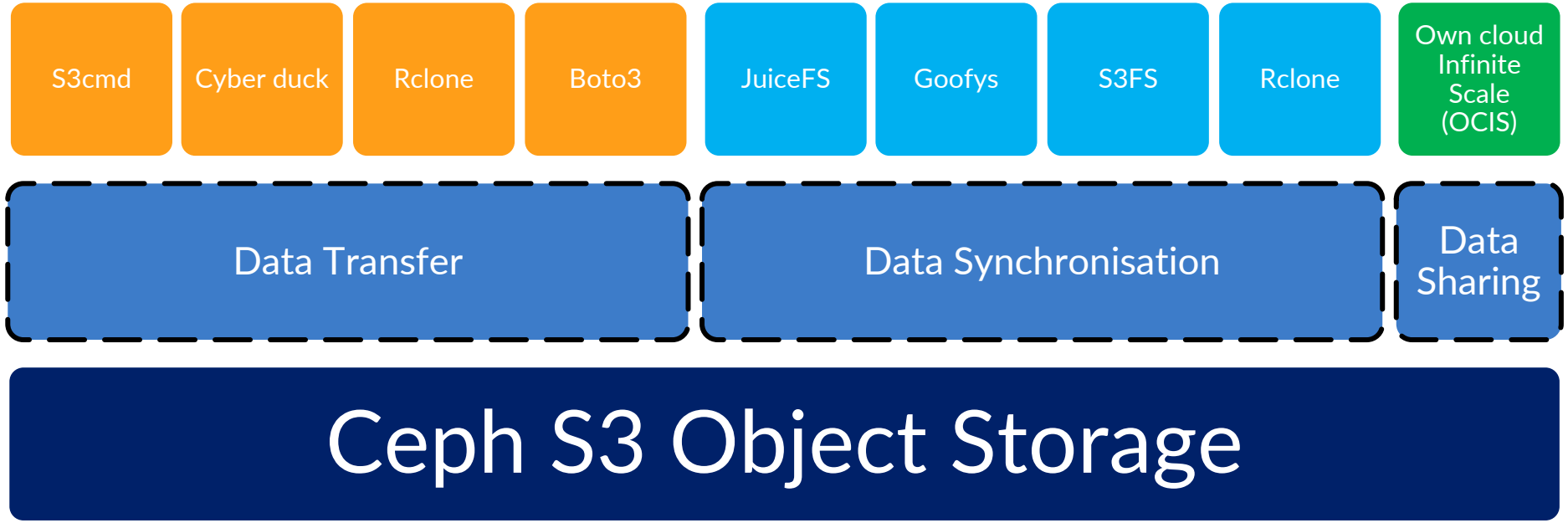
What PSDI do: engage with research communities, promote best practices, offer training, develop services

What is presented today is a part of PSDI service development: we foresee a need in Data Transfer, Synchronisation and Sharing (DTSS) solution

More about PSDI: www.psdi.ac.uk

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Overview of technology trials and use cases of PSDI DTSS Solution

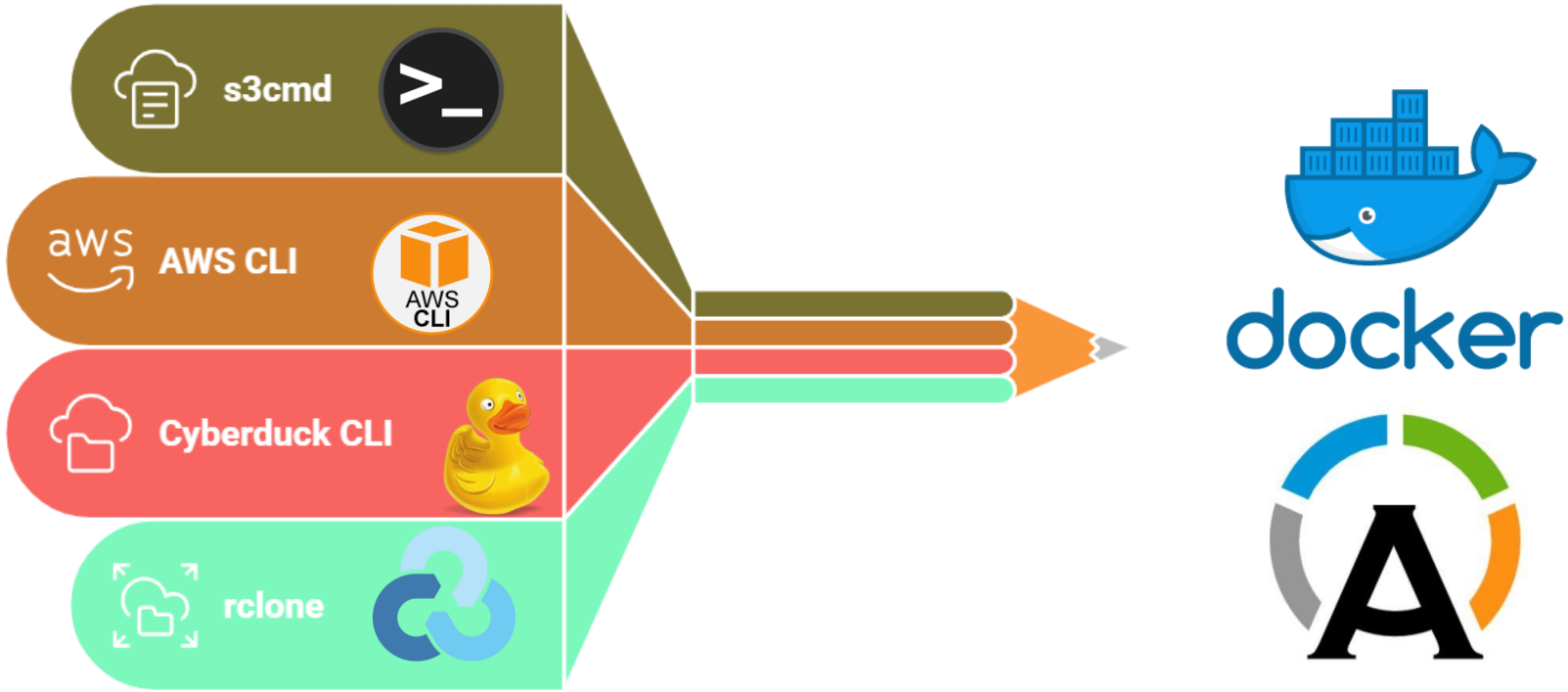


Purpose of Experiments Each tool was assessed to understand its strengths and limitations in handling data workflows, including:

- **Data Transfer**: Efficiently moving large datasets between storage systems.
- **Data Synchronization**: Keeping data in sync across distributed locations or multiple devices.
- **Data Sharing**: Facilitating collaboration and secure data access for users across our network

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Docker & Apptainer containers for data transfer



- **Docker Format**
  - The Docker format is specifically designed for running the data transfer container using the Docker runtime and infrastructure
- **Singularity/Apptainer**
  - Apptainer is particularly well-suited for High-Performance Computing (HPC) environments and shared clusters where many users share the same resources, and security is a primary concern

PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Data Sharing : OCIS

Enabling effective and secure data sharing among teams, departments, and even external partners

Why oCIS?

- Works well with **Ceph Object Storage**.
- Open-source and customizable.
- Supports **scalability & Kubernetes deployment**.

Use Cases:

- e.g., research data management, collaboration.
- Integrations with other PSDI components e.g., Indexing / Computational Notebooks

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Architecture & Deployment

4 Implement Policies

Use Apache Airflow to automate data retention and release policies.

3 Enable Access

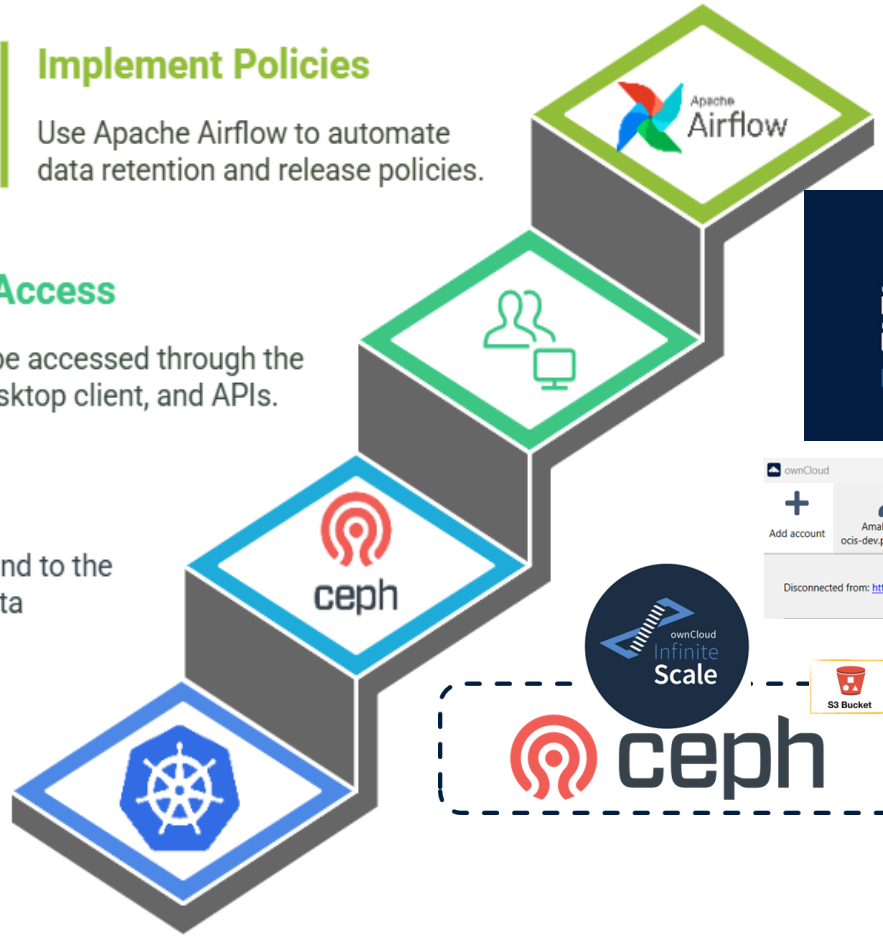
OCIS can be accessed through the web UI, desktop client, and APIs.

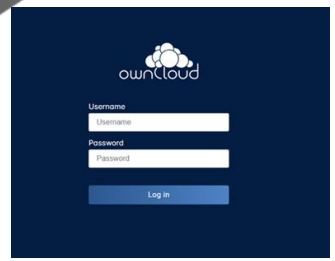
2 Integrate Ceph

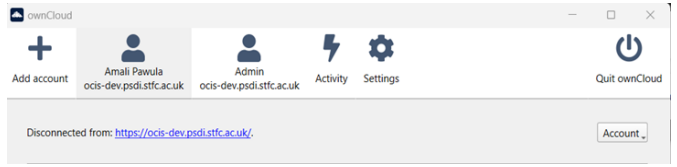
Connect the OCIS backend to the Ceph object store for data management.

1 Deploy OCIS

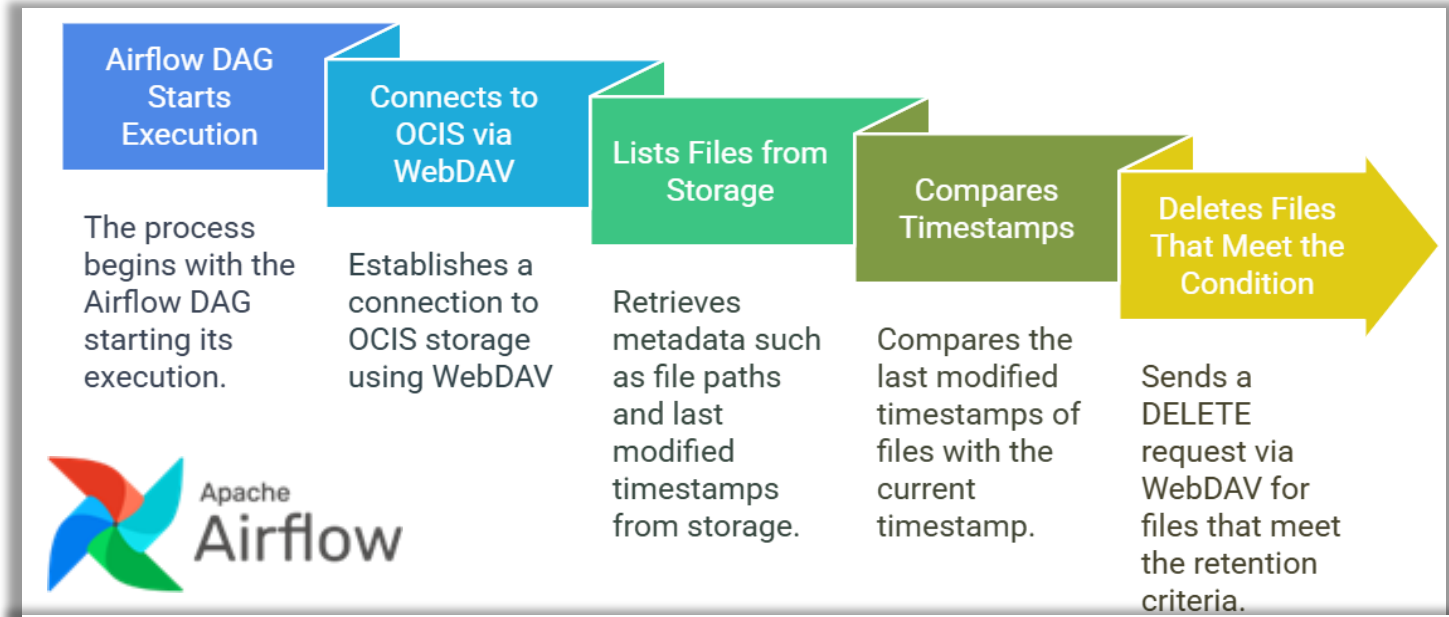
Deploy the OCIS service using Helm charts in the Kubernetes cluster.





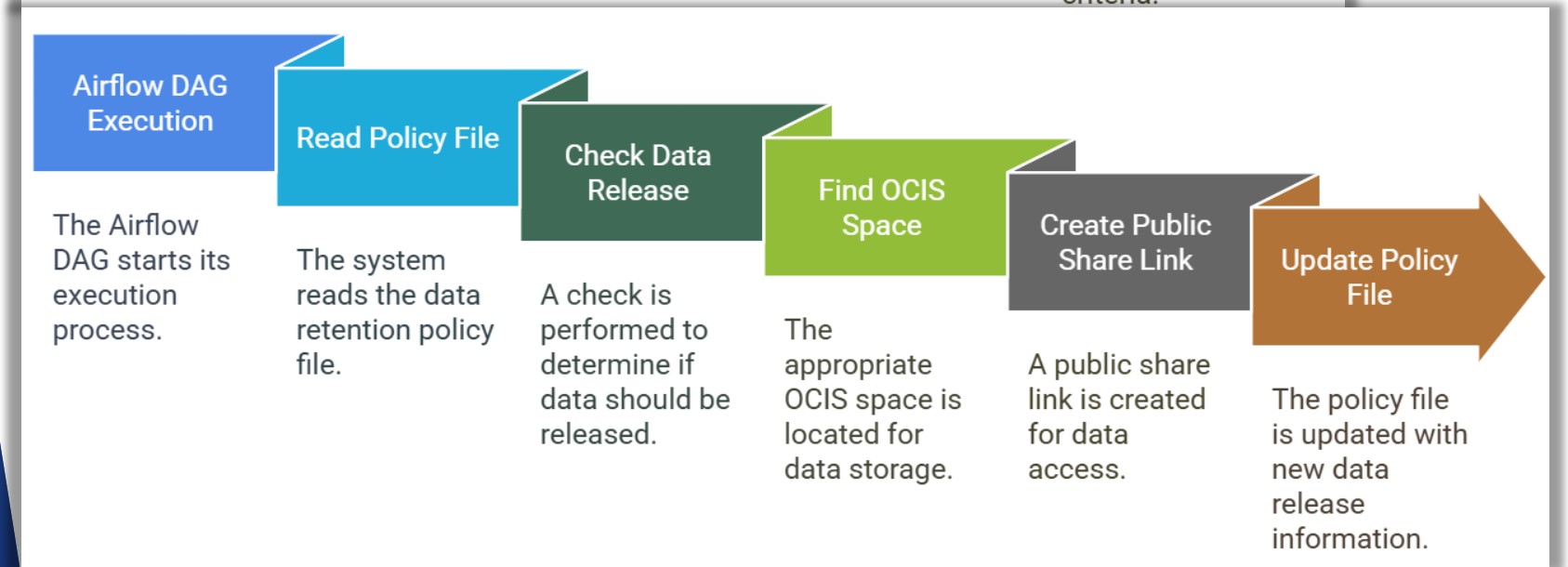


DTSS solution can be combined with workflows solution for defining (and operating) data policies



Data retention policy

- **WebDAV**



Data release policy

- **WebDAV**
- **LibreGraph**

Performance & Challenges

Ref: https://jmeter.apache.org/

Load Testing - Aggregate Report for Login Request with 100 concurrent users – GET HTTP Request

Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Max	Error %	Throughput	Received KB/sec	Sent KB/sec
OCIS Login Request	100	78	70	105	125	169	62	179	0.00%	48.87586	69.16	24.1
TOTAL	100	78	70	105	125	169	62	179	0.00%	48.87586	69.16	24.1

Aggregate Report for File Uploading Request with 100 concurrent users – PUT HTTP Request

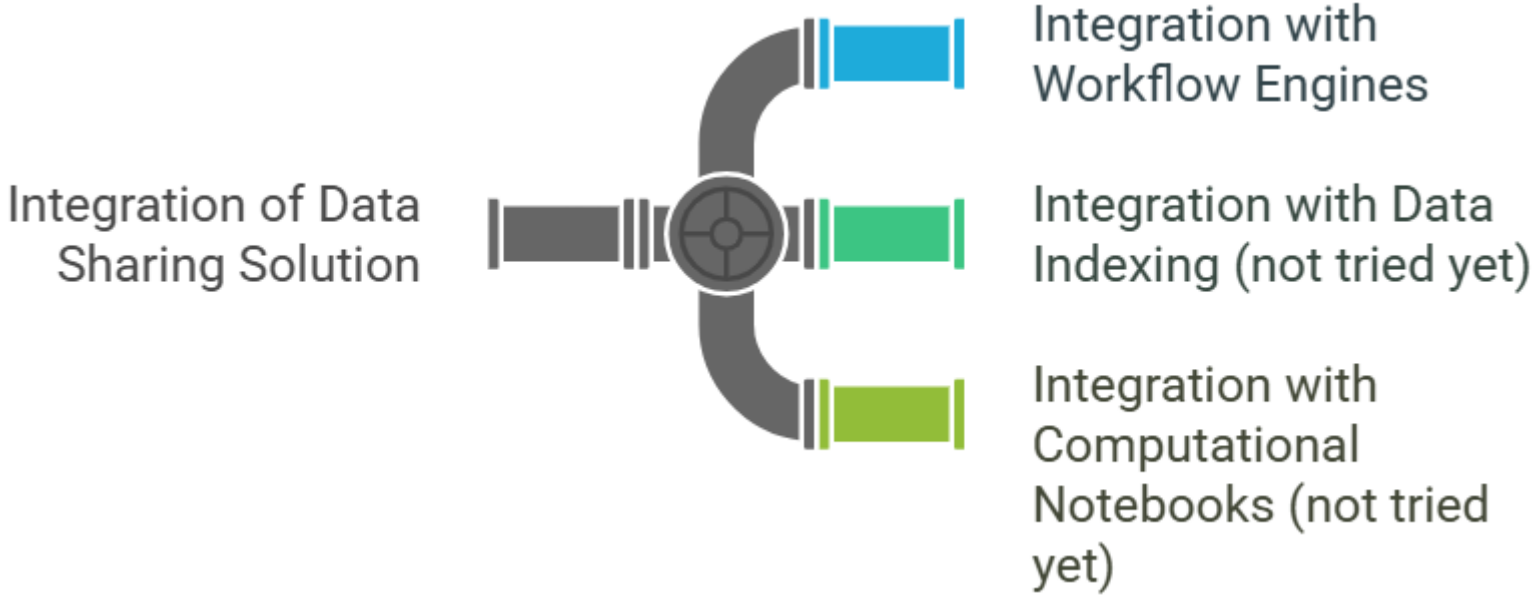
Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Max	Error %	Throughput	Received KB/sec	Sent KB/sec
OCIS File Upload Request	100	290	281	414	441	461	162	463	0.00%	19.33862	12.35	7.88
TOTAL	100	290	281	414	441	461	162	463	0.00%	19.33862	12.35	7.88

Aggregate Report for File Downloading Request with 100 concurrent users – GET HTTP Request

Label	# Samples	Average	Median	90% Line	95% Line	99% Line	Min	Max	Error %	Throughput	Received KB/sec	Sent KB/sec
OCIS File Download Request	100	307	289	425	508	582	168	595	0.00%	30.75977	29.83	9.22
TOTAL	100	307	289	425	508	582	168	595	0.00%	30.75977	29.83	9.22

Conclusion and next steps

- Containerised DTSS tools offered on top of Ceph storage can be an interim DTSS solution for our project, especially with newer versions of Ceph that support multitenancy *)
- The adoption of a feature-rich data sharing solution like OCIS or OpenCloud will be of course advantageous, and we will keep looking into it
- Integrations of a data sharing solution with other components will be of interest:



*) See https://ceph.io/en/news/blog/2025/enhancing-ceph-multitenancy-with-iam-accounts/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Thank you