



OBJECT STORAGE @ MPCDF

- USAGE AND USE-CASES

John Alan Kennedy
Cloud Enabling Team

WHAT IS MPCDF OBJECT STORAGE



- Scalable object storage services, compatible* with the Amazon S3 protocol.
- Solutions for Projects / Individual users
- Globally available
- Policies
 - Data Access
 - Data Lifecycle
 - Versioning
- Temp URLs
 - Time limited URL for download/upload

* Full coverage of the AWS S3 API functionality is not guaranteed

OBJECT STORAGE (PROJECTS)

Large scale data transfer and sharing (For Projects 10s-100s TiBs)

- Object Storage (2020*)

- S3 compatible API

- *PUT/GET data*
- *Policies (life-cycles, versioning etc)*
- *Sharing (temp urls)*

- Globus on roadmap

- 11PiB Total storage

- Project based access (Rental)

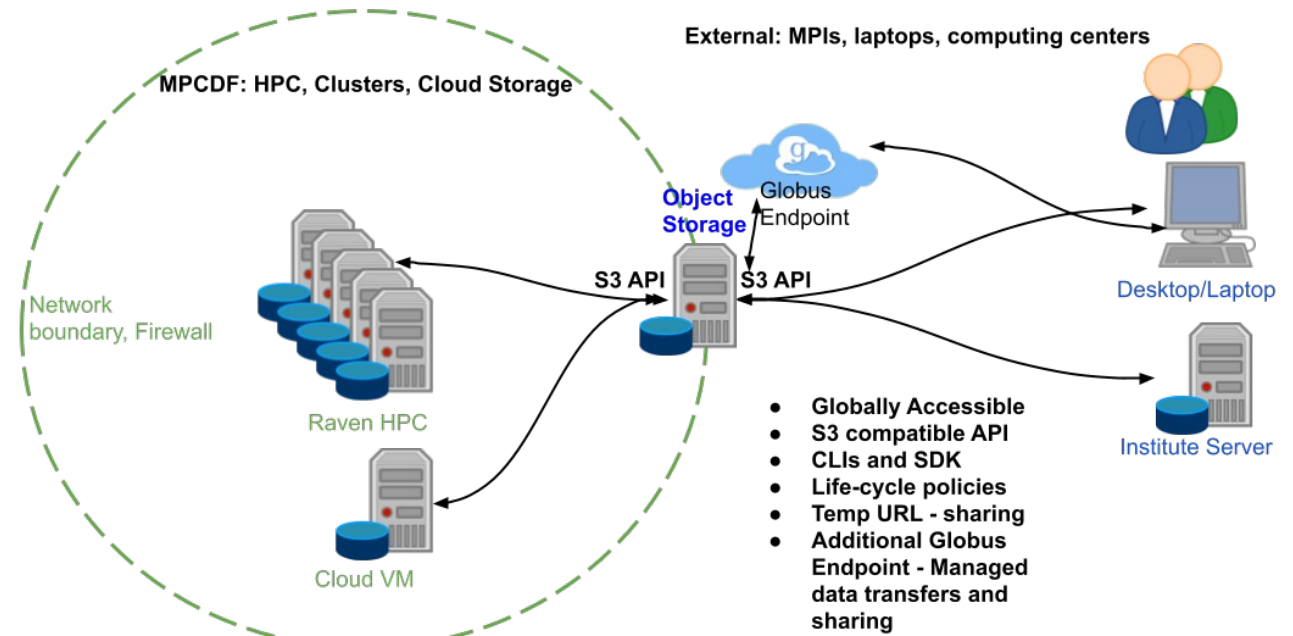
- Use-Cases

- Services with native S3 backend

- Data Analytics (DVC, DataLad)

- Global Project Data Store

- Share/Publish Datasets



Global Access:

- Services in MPCDF and external

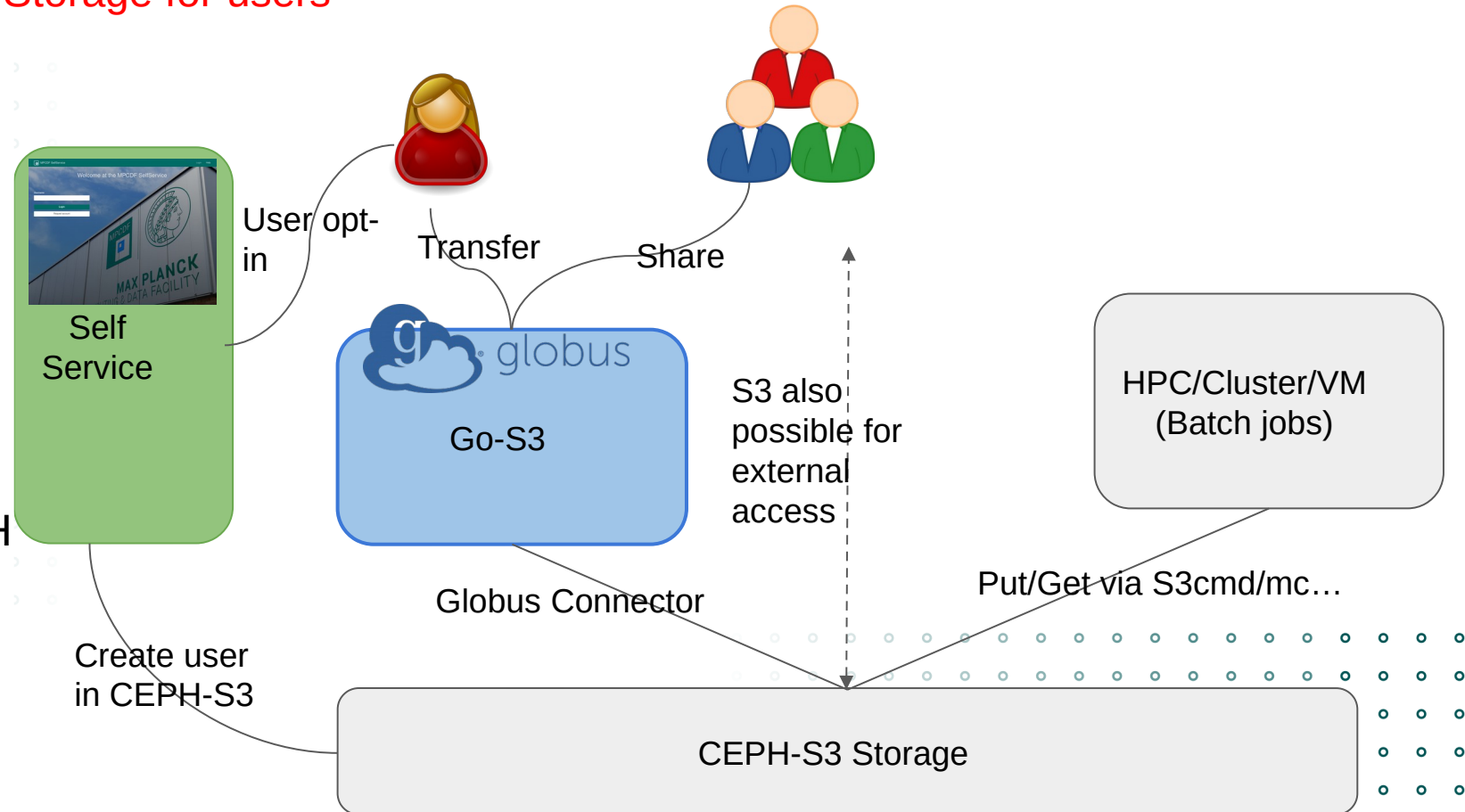
- Many services and tools

* Available since 2020
Extension in 2023

NEXUS-S3 (USERS)

Permanent S3 / Object Storage for users

- Nexus-S3 (soon...)
 - Object Storage for users
 - S3 API
 - Globus Enabled
 - 1TiB free data
- Workflow
 - User self-service opt-in
 - Account created in CEPH
 - S3 from batch and external
 - Data Transfer/Share via Globus



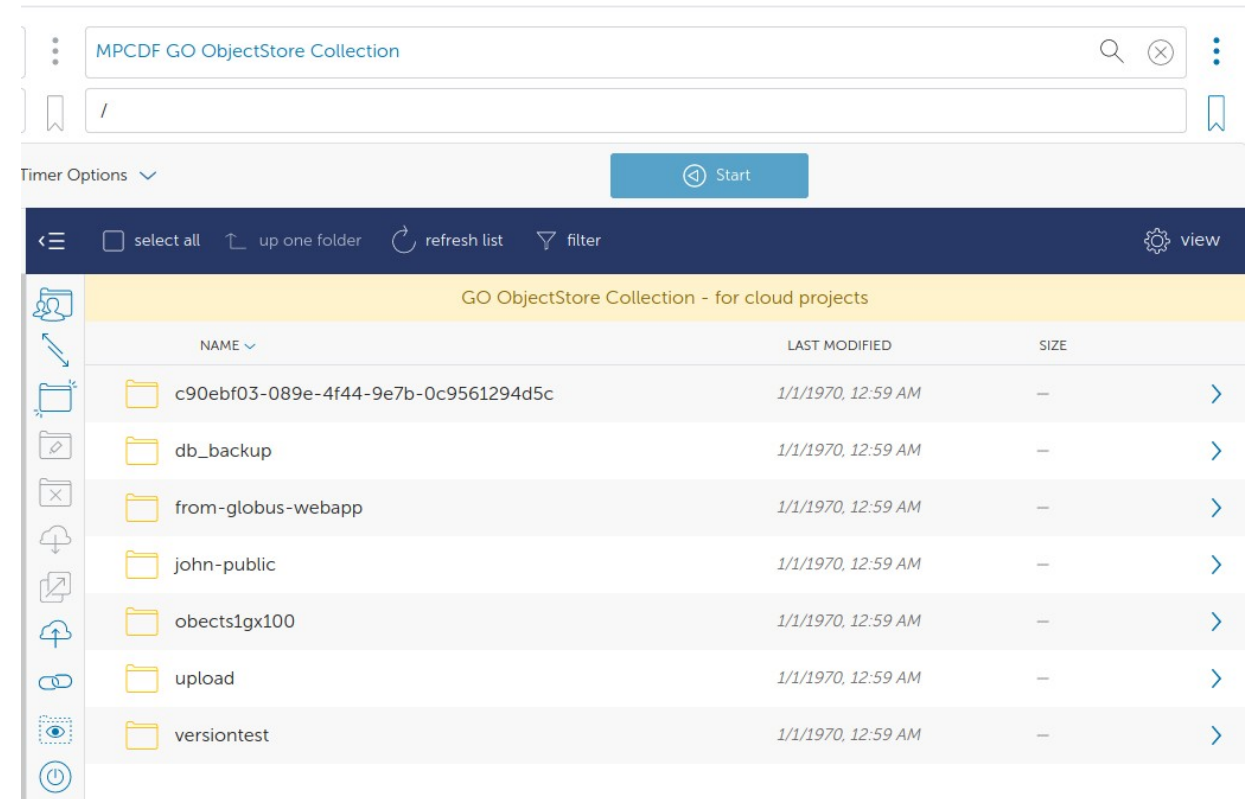
S3 and Globus access same storage = Best of both worlds

HOW TO ACCESS OBJECT STORAGE

- A range of CLI tools
 - Rclone, s3cmd, minio
 - Rclone, minio modules available
- API
 - SDKs in go, python, ruby..
 - Boto3 python library
- Globus

 MPCDF GO Nexus S3 Collection
Subscribed Mapped Collection (GCS) on MPCDF GO-S3

 MPCDF GO ObjectStore Collection
Subscribed Mapped Collection (GCS) on MPCDF GO-S3



Access as standard Globus collection

USE-CASES



- **Data Lake**

- Collect data from batch jobs, or external systems

- **Data Sharing**

- Public bucket
- Globus

- **Storage backend for service**

- Data Upload/Download
- Cache.

- **Seen policies used for**

- Restricting access to IP ranges
- Tag individual files for public access

- **Seen temp urls:**

- Temp upload and download

USE-CASES



- **Data Lake**

- Collect data from batch jobs, or external systems

- **Data Sharing**

- Public bucket
- Globus

- **Storage backend for service**

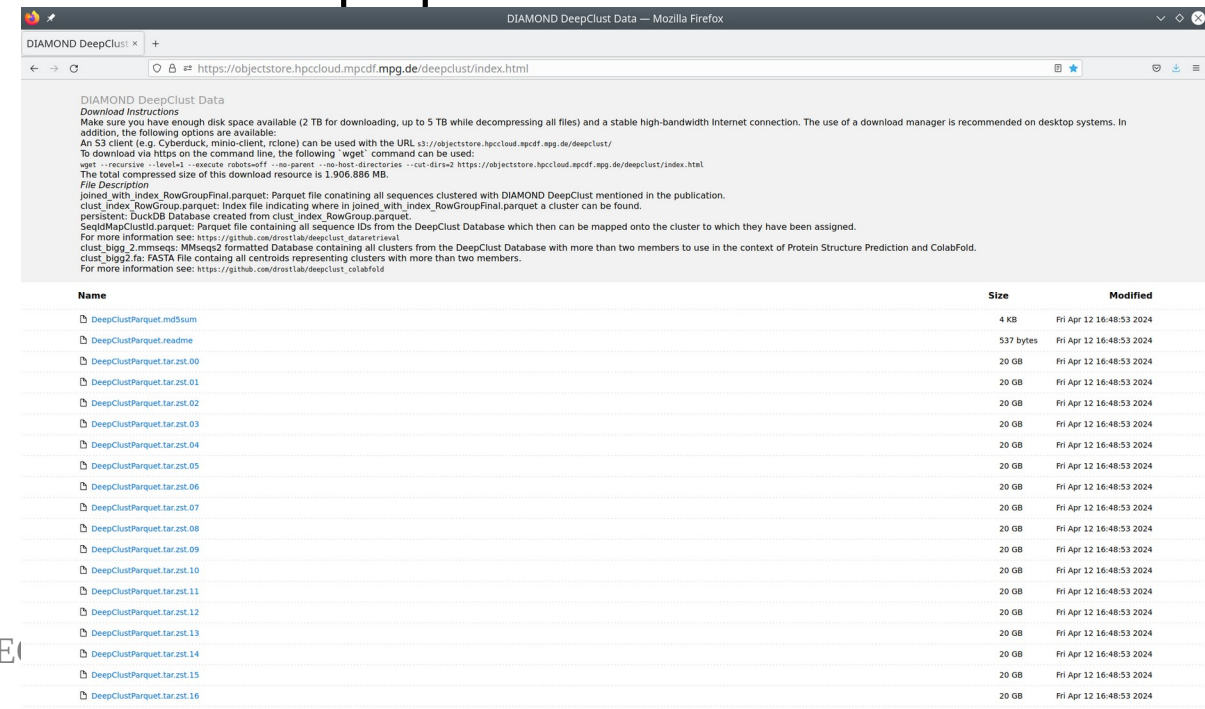
- Data Upload/Download
- Cache.

- **Seen policies used for**

- Restricting access to IP ranges
- Tag individual files for public access

- **Seen temp urls:**

- Temp upload and download



DIAMOND DeepClust Data

Download Instructions
Make sure you have enough disk space available (2 TB for downloading, up to 5 TB while decompressing all files) and a stable high-bandwidth Internet connection. The use of a download manager is recommended on desktop systems. In addition, the following options are available:
An S3 client (e.g. Cyberduck, minio-client, rclone) can be used with the URL `s3://objectstore.hpcloud.mpcdf.mpg.de/deepclust/`
To download via https on the command line, the following `wget` command can be used:
`wget --recursive --level=1 --execute robots-off --no-parent --no-host-directories --cut-dirs=2 https://objectstore.hpcloud.mpcdf.mpg.de/deepclust/index.html`
The total compressed size of this download resource is 1.906.886 MB.

File Description
joined_with_index_RowGroupFinal.parquet: Parquet file containing all sequences clustered with DIAMOND DeepClust mentioned in the publication.
clust_index_RowGroup.parquet: index file indicating where in joined_with_index_RowGroupFinal.parquet a cluster can be found.
persistant_DuckDB Database created from clust_index_RowGroup.parquet.
SeqMapClustId.parquet: Parquet file containing all sequence IDs from the DeepClust Database which then can be mapped onto the cluster to which they have been assigned.
For more information see: https://github.com/drostlab/deepclust_dataretrieval
clust_bigg_2_mmsseqs: MMseqs2 formatted Database containing all clusters from the DeepClust Database with more than two members to use in the context of Protein Structure Prediction and ColabFold.
clust_bigg2.fasta: FASTA File containing all centroids representing clusters with more than two members.
For more information see: https://github.com/drostlab/deepclust_colabfold

Name	Size	Modified
DeepClustParquet.md5sum	4 KB	Fri Apr 12 16:48:53 2024
DeepClustParquet.readme	537 bytes	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.00	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.01	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.02	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.03	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.04	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.05	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.06	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.07	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.08	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.09	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.10	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.11	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.12	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.13	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.14	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.15	20 GB	Fri Apr 12 16:48:53 2024
DeepClustParquet.tar.zst.16	20 GB	Fri Apr 12 16:48:53 2024

Data Publishing →
Hajk-Georg Drost, Benjamin Buchfink
(Diamond, MPI for Biology Tübingen)
Klaus Reuter (MPCFD)

MANY THANKS TO



- Robert Hish
- Florian Kaiser
- Michele Compostella
- Tom Zastrow
- Nicolas Fabas

- Frank Berghaus
- Maximiliano Geier
- Brian Standley



THANK YOU – QUESTIONS WELCOME!