Integrated, heterogeneous data access in INFN-Cloud and beyond

Stefano Stilio
stefano.stilio@lnqs.infn.it
on behalf of the INFN Cloud Project
cloud@infn.it

EGI Conference 2020
2-5 November 2020 - Online
Overview

- The INFN Cloud
- Object Storage @INFN Cloud
- Object Storage integration with the INFN Cloud PaaS
  - use cases
- Object Storage integration with high level storage solutions
  - examples and a use case
- Work in progress
The INFN Cloud

• **INFN Cloud** is a federation of existing infrastructures
  • the **INFN Cloud backbone**: it consists of two tightly coupled federated sites (BARI and CNAF)
  • a scalable **set of satellite sites**, geographically distributed across Italy, and loosely coupled.

• **INFN Cloud aims at offering a full set of high-level cloud services to INFN user communities**
  • the service catalogue is not static: new applications are included through a defined “on-boarding” process for new use-cases
The INFN Cloud

• **Key enabling factors for the federation** are:
  
  • leverage the same authentication/authorization layer based on INDIGO-IAM;
  
  • agree on a consistent set of policies and participation rules (user management, SLA, security, etc.);
  
  • transparent and dynamic orchestration of the resources across all the federated infrastructures through the INDIGO PaaS Orchestrator
The INFN Cloud Object Storage

Besides the IaaS and Paas services, INFN Cloud offers to its end users an **object storage service based on OpenStack Swift**, whose main characteristics are:

- Distributed set-up: data is replicated over two different data centers about 600km away from each other
- Swift and S3 APIs
- Different storage policies for different data sets
- Keystone-based authentication
- Indigo IAM on top of Keystone
The INFN Cloud Object Storage

The INFN Cloud Object Storage service is the main data backend for end user applications, as it ensures

- resilience
- high availability
- easy access
- authenticated access
- ubiquitous access
The INFN Cloud Object Storage

INFN Cloud Object storage is also used internally by the backbone as:

- OpenStack image repository
- OpenStack image snapshot storage backend
- OpenStack block device backup
- backup for DB data, configuration data, ….
The INFN Cloud Object Storage

• End users can access the INFN Cloud Object Storage “as is”, exploiting directly the OpenStack Swift and S3 APIs.

• High level services offered by the INFN Cloud PaaS give users an opportunity to take advantage of this service without needing to deal with the details of low level interaction with the storage APIs and allowing them to focus on their research goals.
Integration with the INFN Cloud PaaS

The main goal of the INFN Cloud PaaS is that of providing to the INFN scientific communities easy to use, high level instruments that can make the researchers work easier when it comes to interacting with computing resources.

As data is the “core business” of scientific institutions like INFN, making access to data easier is of paramount importance for the INFN Cloud project.
Use case:
Object storage for experimental data

The **Cygno** Dark Matter experiment uses the INFN Cloud Object Storage as its main back-end for DAQ data, and fully integrates the OpenStack Swift APIs in its data analysis framework. In this way, analysis can run anywhere, without having to worry about data retrieval.

The whole **Cygno** data analysis framework has been ported to the INFN Cloud. A pre-configured VM holding all the needed software and libraries and exposing both Jupyter and ssh can be started with just a couple of clicks.
Use case:
Jupyter with persistent storage

The INFN Cloud PaaS allows users to select an S3 bucket to be mounted as a “posix” file system in a jupyter notebook.

This is as easy as filling four fields in the description of the deployment.
Integration with other software

Integration with high level tools and ready-to-go solutions improves the usability of object storage and the overall user experience.

Compatibility with a number of open source software tools has been tested and validated.

A number of these has been chosen as building blocks for high level storage services that have been/are being integrated in the INFN Cloud PaaS.
Integration with other software

Among many tools that can be integrated with an object storage service exposing Swift and S3 APIs, the following are presently being used or tested in different INFN Cloud use cases:

- Duplicati
- ownCloud
- Rclone
- Minio

Other tools have been tested, and we may integrate them in future use cases:

- Duplicity, Nextcloud, S3cmd, awscli, s3fs, goofys,....
Example: ownCloud/Nextcloud

ownCloud and Nextcloud can be used as user interfaces for the storage of scientific data

- User friendly WEB UI
- Object Storage as main storage back-end very efficient as metadata is kept on the application DBMS
- full control on the user side on data access, ACL, sharing
- programmatic access via webdav clients (e.g Rclone) for scientific workloads
Example: Rclone

used to mount OS buckets as posix folders, syncing remote buckets locally and viceversa, retrieving and uploading single files

as a webdav client for ownCloud/Nextcloud it allows for programmatic access to data, maintaining the authn/authz granularity offered by ownCloud/Nextcloud
Example: Duplicati

Duplicati is a backup software to store encrypted backups online.

It has been integrated in some of the INFN Cloud services for periodically backing up databases and configuration data.

End users only need to perform a minimal setup via the duplicati web GUI in order to have the service running.
Use case: “Cloud Storage Service”

As an example, the INFN Cloud “Cloud Storage Service” uses

- ownCloud as an interactive frontend to S3 storage
- Rclone for programmatic access
- Duplicati for the backup of DB and configuration data
- Openstack Swift as a data backend.
Use case: “Cloud Storage Service”

- Developed to meet the requirements of a specific scientific collaboration...
- ...and to satisfy similar requirements from other collaborations
- Easily deployed by the end user by filling up a simple form
- Also integrates status monitoring
Work in progress: Minio GW + OPA + IAM

Minio gateway can be used as a front end for the INFN Cloud object storage, and this gives:

- Easy to use web interface
- command line client
- S3 APIs

But the fact that Minio has limited access control capabilities can be a showstopper for some use cases
Work in progress:
Minio GW + OPA + IAM

A possible solution to overcome the limited access control capabilities of Minio Gateway is that of integrating it with Open Policy Agent

https://www.openpolicyagent.org/

Indigo IAM authn

MINIO Gateway

single “service” credentials

OpenStack Swift

OPA authz
Work in progress: Minio GW + OPA + IAM

OPA can take care of managing granular access policies to different buckets relieving all the complexity of authentication and authorization from the storage backend.
Conclusions

• INFN Cloud offers a solid and resilient Object Storage service

• Our goal is making data management and access easier for our user communities

• The way we try to do this is via the integration with high level services that users can approach with no knowledge of low level details
Thanks for your attention!!!