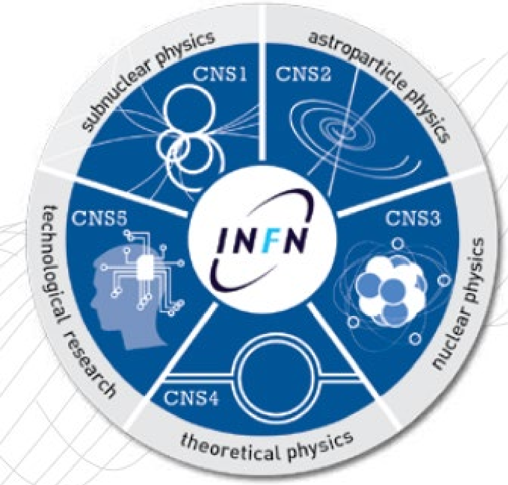
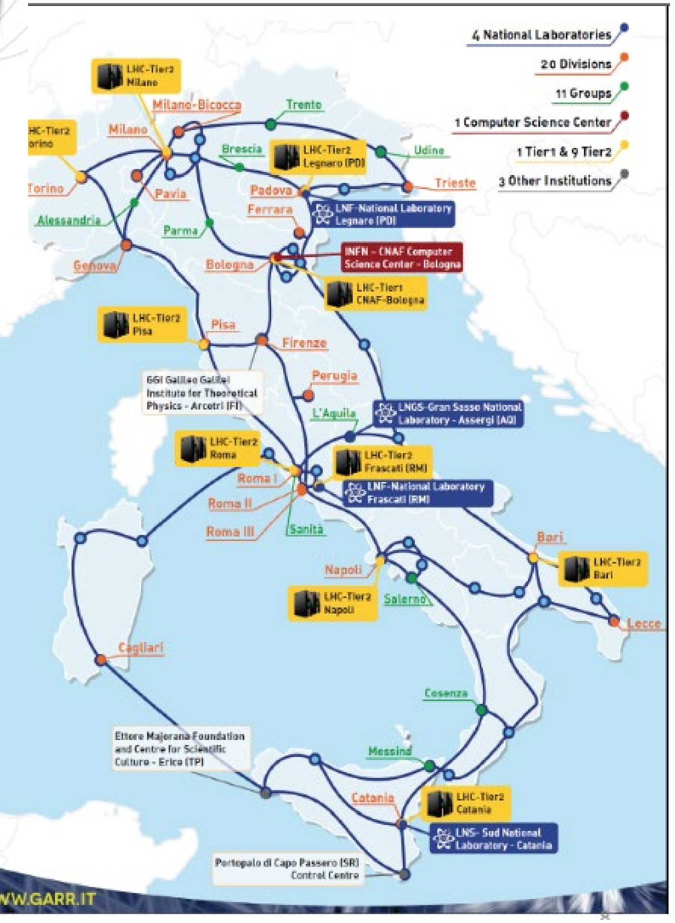
The S3 Object Storage Service on INFN Cloud

Ahmad Alkhansa (ahmad.alkhansa@cnaf.infn.it)
Diego Ciangottini (ciangottini@pg.infn.it)
Alessandro Costantini (alessandro.costantini@cnaf.infn.it)
Federico Fornari (federico.fornari@cnaf.infn.it)
Jacopo Gasparetto (jacopo.gasparetto@cnaf.infn.it)
Giada Malatesta (giada.malatesta@cnaf.infn.it)
Diego Michelotto (diego.michelotto@cnaf.infn.it)
Massimo Sgaravatto(massimo.sgaravatto@pd.infn.it)
Daniele Spiga (daniele.spiga@pg.infn.it)
Stefano Stalio (stefano.stalio@lngs.infn.it)

Italian Institute for Nuclear Physics INFN

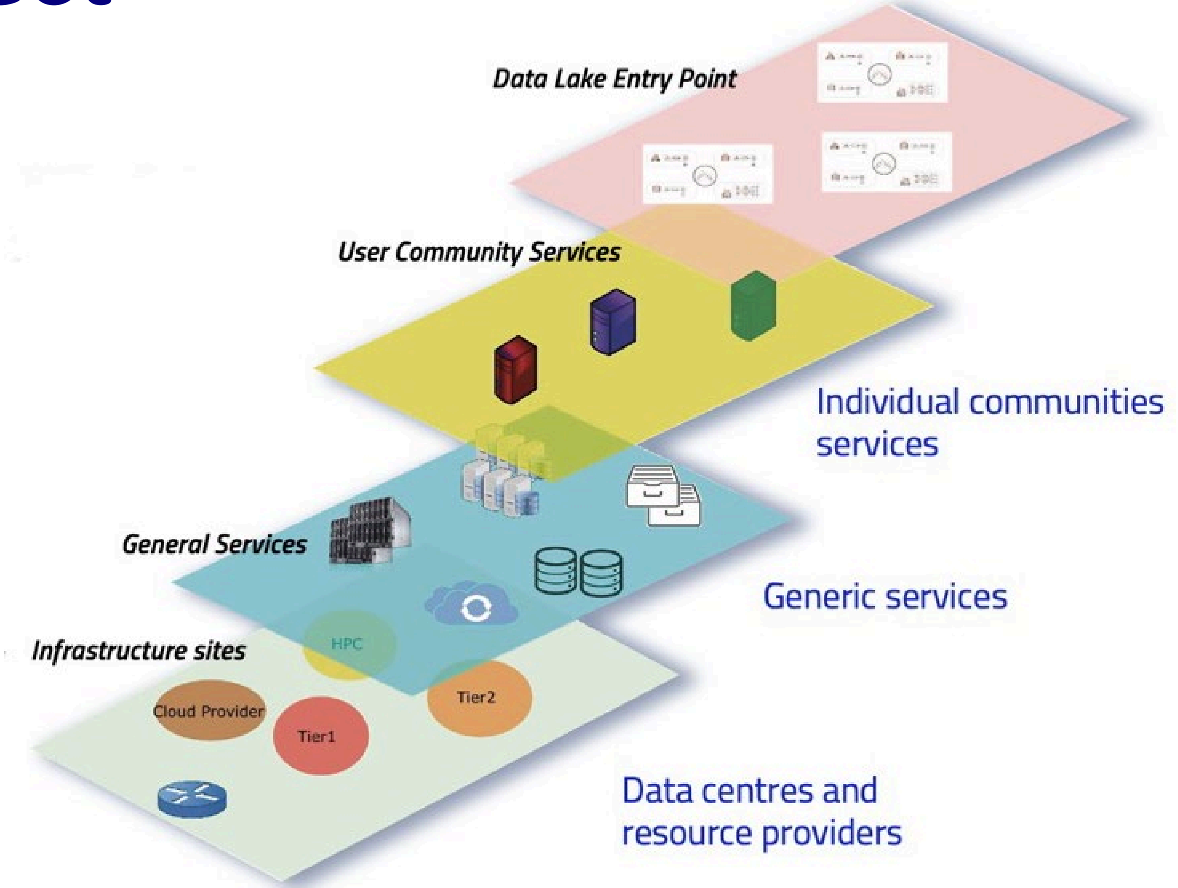
- 5 lines of research
  - With computing as a transversal needs
- Facilities
  - 4 national laboratories
  - 20 divisions
  - 6 associated groups
  - 3 national centers and schools
  - 1 international consortia
- Strong participation on national and international projects and collaborations





The INFN DataCloud Project

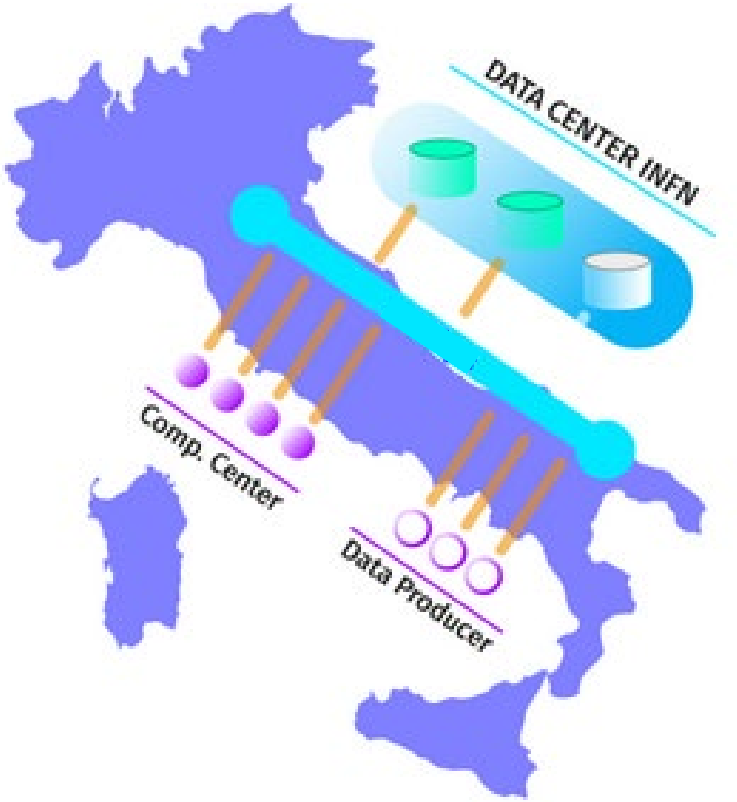
- The DataCloud Project manages all **core activities related to computing @INFN and its projects**
  - Development, implementation & management of the INFN Datalake architecture
  - Development of ISO-Certified solutions mainly for clinical and omics data management
  - Support to users and to the management and operation of all INFN sites (both Grid and Cloud paradigms)
  - Development of new services
- Focus on **Integration of resources**, methods, people, solutions
- Modular architecture based on **service composition**
- The **INFN foundation** of all the NRRP computing-related initiatives



Proposed architecture for ICSC and TeRABIT projects

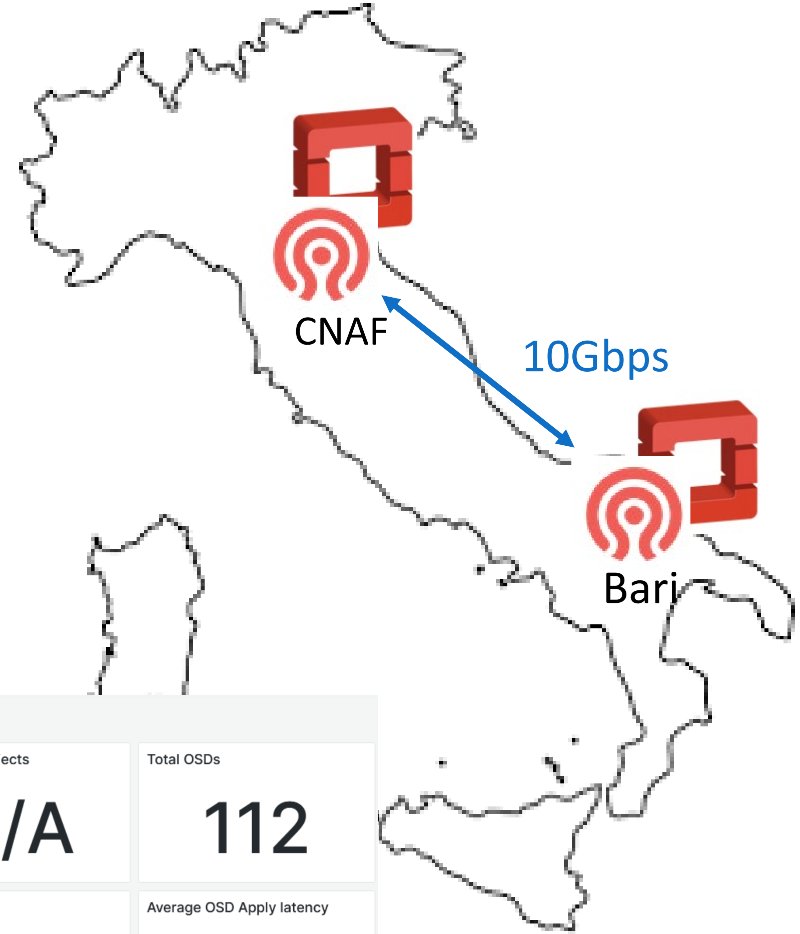
Context and Use Case

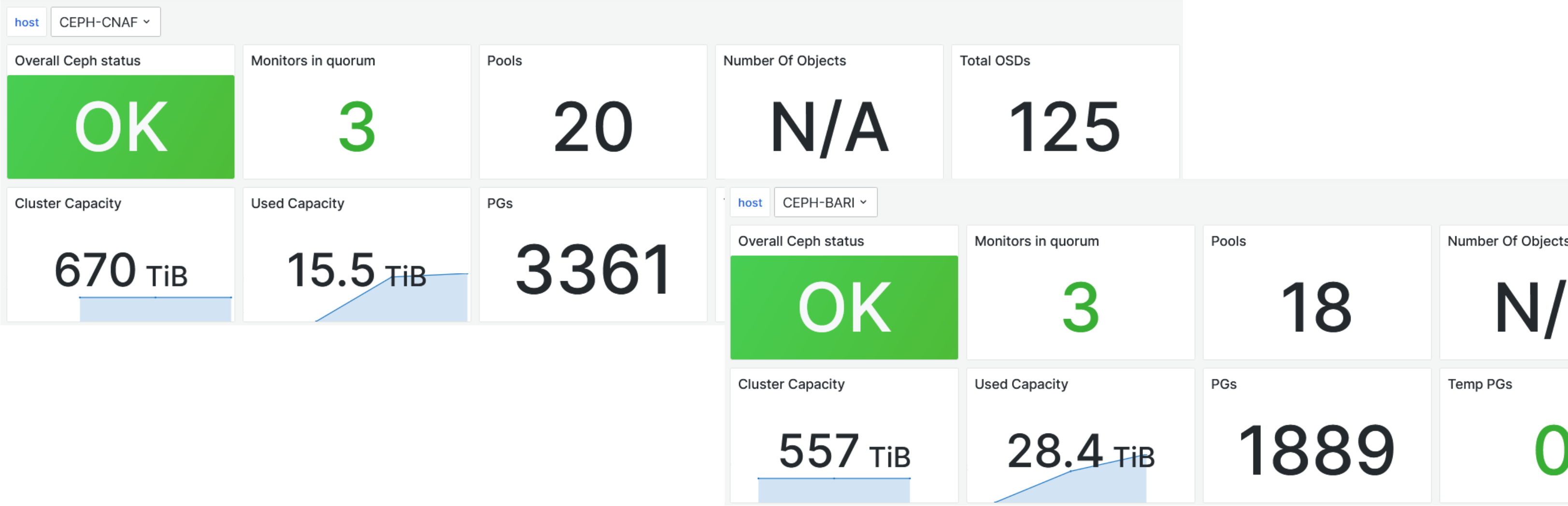
- **MinIO Gateway** is **not supported** anymore
  - Still in used today
- **Distribution of Data** with multiple access points
- Technology that allows **scalability of resources**
- **Federated Authentication** with **fine-grained authorization**
- Allow users to access **Cloud storage** in **POSIX**-like manner



INFN Cloud distributed services

- INFN Cloud rely on 2 geographical distributed sites (**BackBone**) hosting the Cloud core services
  - OpenStack deployment
  - **CEPH cluster**
  - Other ancillary services





host CEPH-CNAF

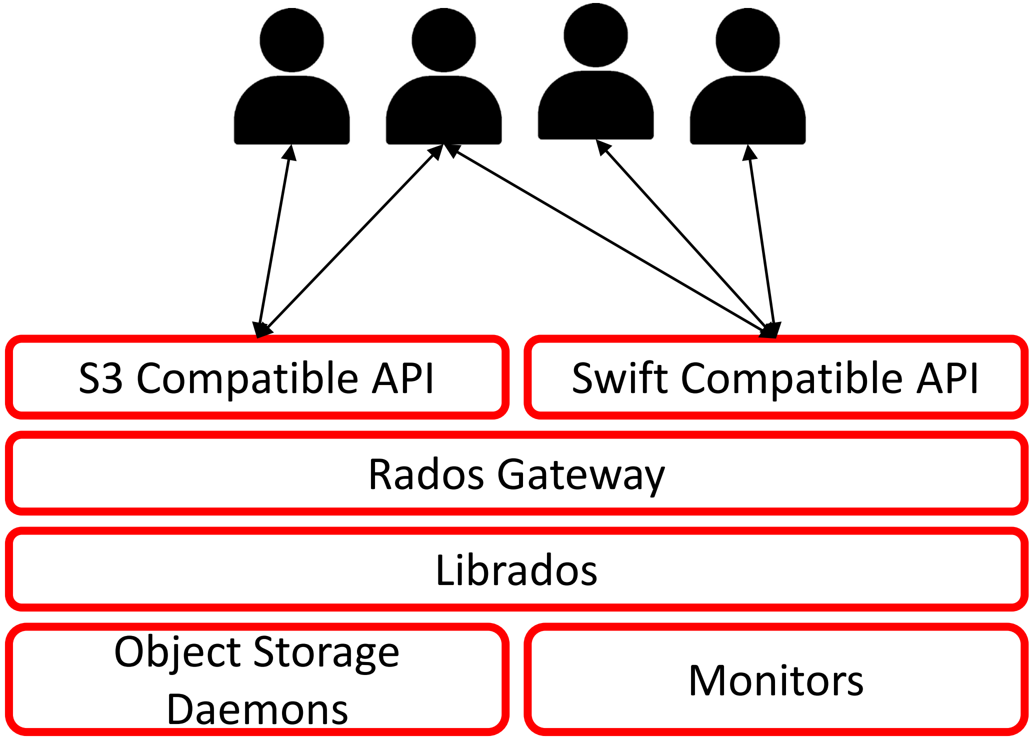
Overall Ceph status	Monitors in quorum	Pools	Number Of Objects	Total OSDs
OK	3	20	N/A	125
Cluster Capacity	**Used Capacity**	**PGs**		
670 TiB	15.5 TiB	3361		

host CEPH-BARI

Overall Ceph status	Monitors in quorum	Pools	Number Of Objects	Total OSDs
OK	3	18	N/A	112
Cluster Capacity	**Used Capacity**	**PGs**	**Temp PGs**	**Average OSD Apply latency**
557 TiB	28.4 TiB	1889	0	24.3

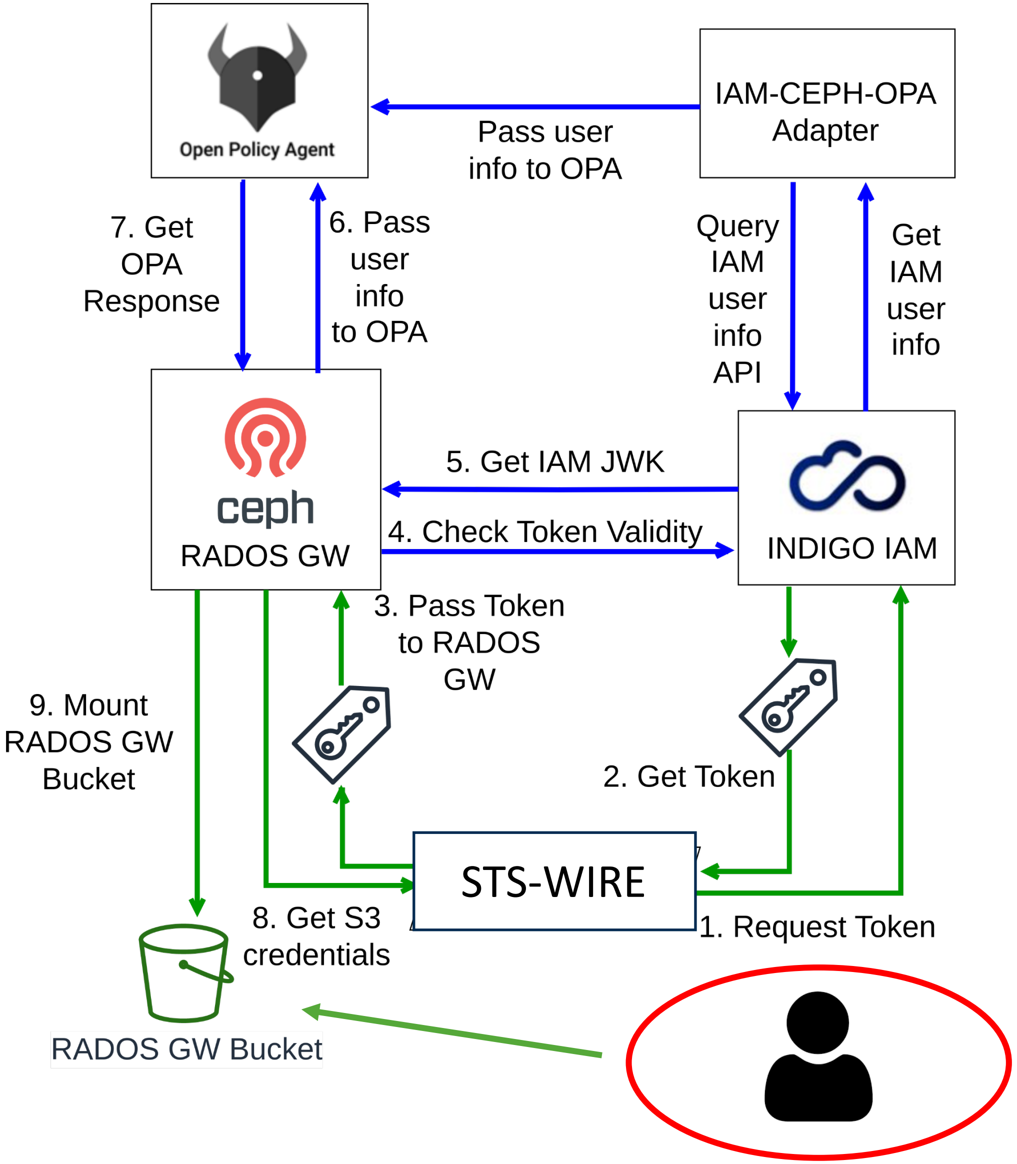
Ceph Object Store

- **Rados Gateway** (RGW) is a ceph object storage **interface**.
- Supports **Secure Token Service** (STS) operations.
- Allows the addition of **OpenID Connect providers**.
- Integrates with **Open Policy Agent** (OPA) for **fine-grained authorization.**



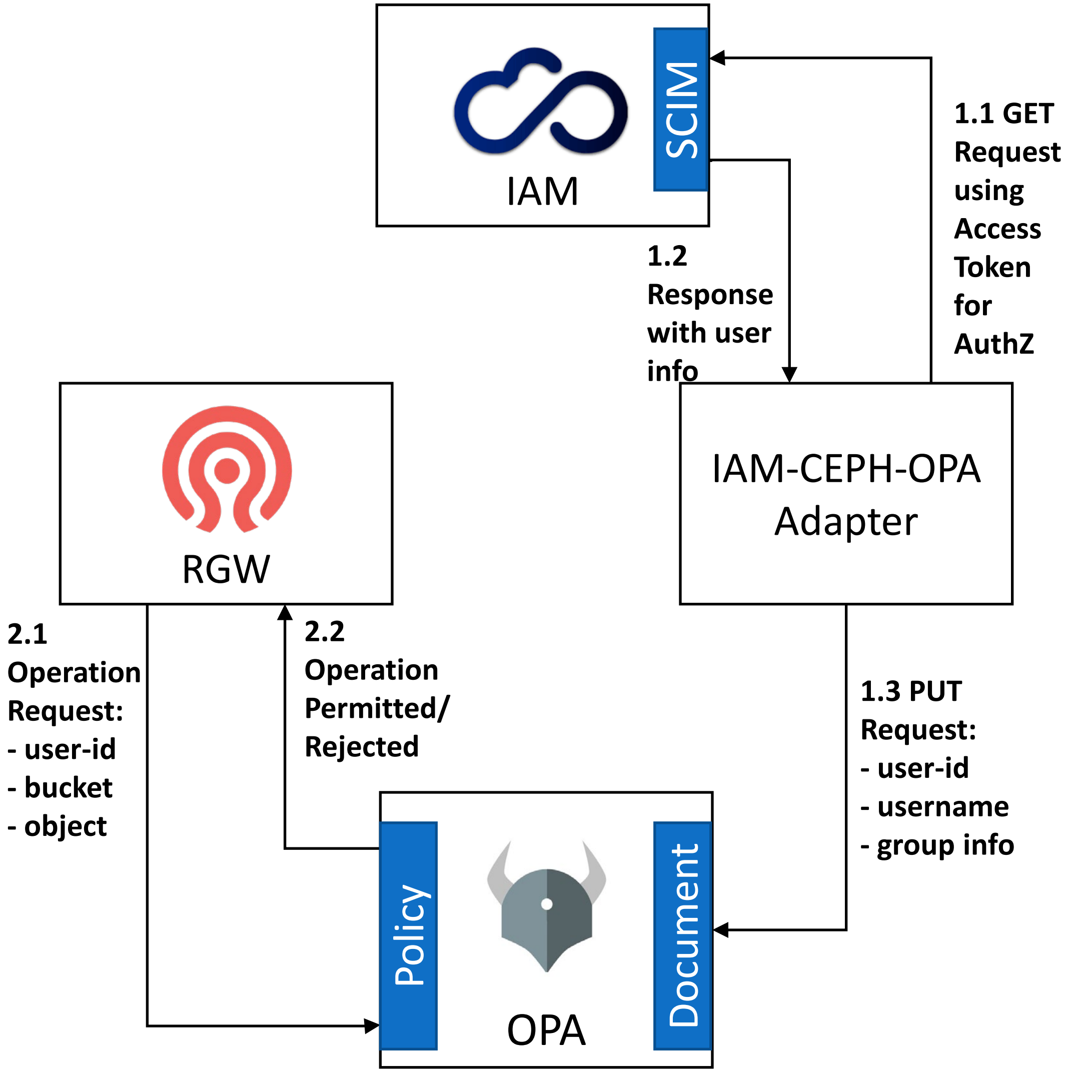
Service integration

- **STS-Wire** is a wrapper of **Rclone** including the IAM **AuthN/AuthZ** configuration.
- The library retrieves IAM access token for performing **STS with RGW**.
- **RGW** validates the token with IAM then sends an **authorization request to OPA.**
- **OPA's response** depends on the content of **RGW input**, existing **policies** and **information** received from the **adapter**.



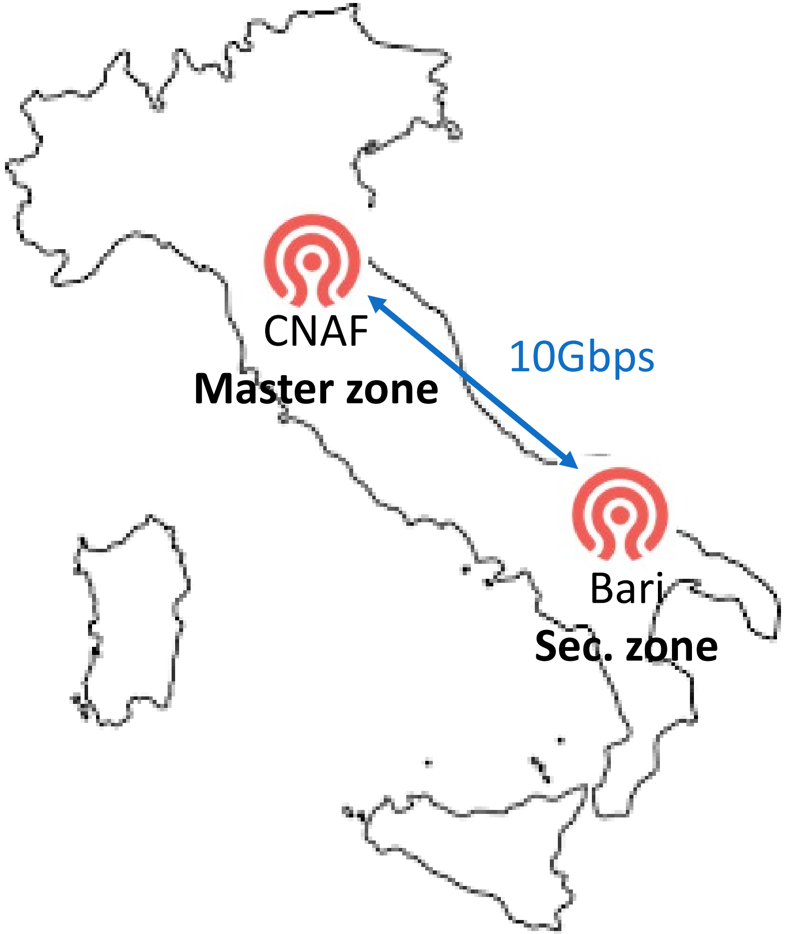
More about IAM-CEPH-OPA Integration

- RGW request to OPA contains only **Token subject claim** (user-id).
- Written **Policies** allow the creation of **buckets** called with IAM usernames.
- The adapter **takes advantage** of System for Cross-domain Identity Management (**SCIM**).
- The adapter interacts with the REST API of OPA to **upload** the necessary data.
- OPA performs queries to **map token subject to username**.



CEPH MULTI-SITE for object replication

- **Multi-zone approach**
  - **Master** (CNAF) and **Secondary** (Bari) zone configuration
  - 1 REALM
  - **Active-Passive** configuration
  - Can be easily switched (manual configuration)
- 3 **RGW instances** on each zone
- 2 **HAProxy** acting as LB and traffic shaping

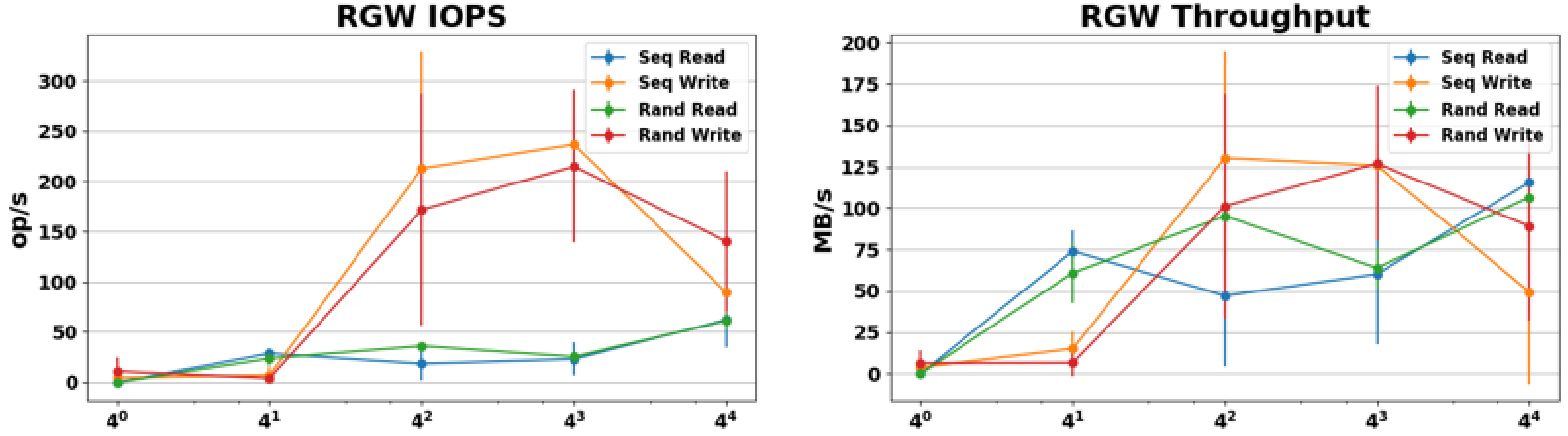


Performance test

- A variable number of (power of 4) parallel clients mounting a bucket
  - FIO with 1 GB file andblocksize of 4MB and 4KB
  - Using rclone+stswire
  - Compared with S3FS+plugin developed for IAM
- Output
  - Server side (from CEPH monitoring, Prometheous enabledon CEPH MGR)
    - **IOPS** and **Throughput**
  - Client side (from FIO)
    - **IOPS** and **Troughput**

Performance test: blocksize=4 MB

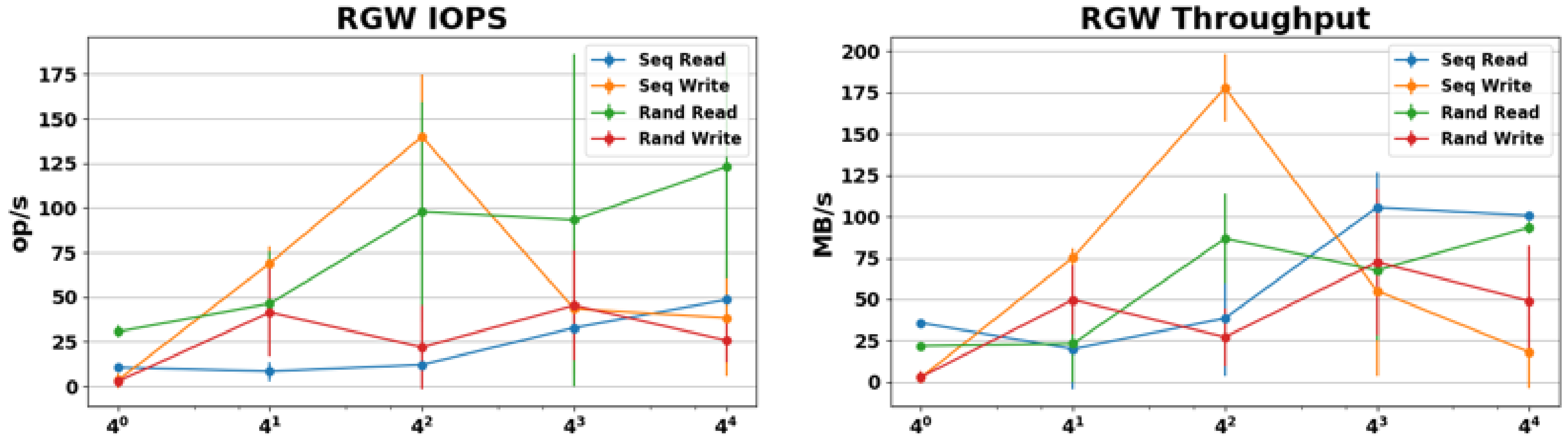
sts-wire performance - bs = 4M, fs = 1G



RGW IOPS

RGW Throughput

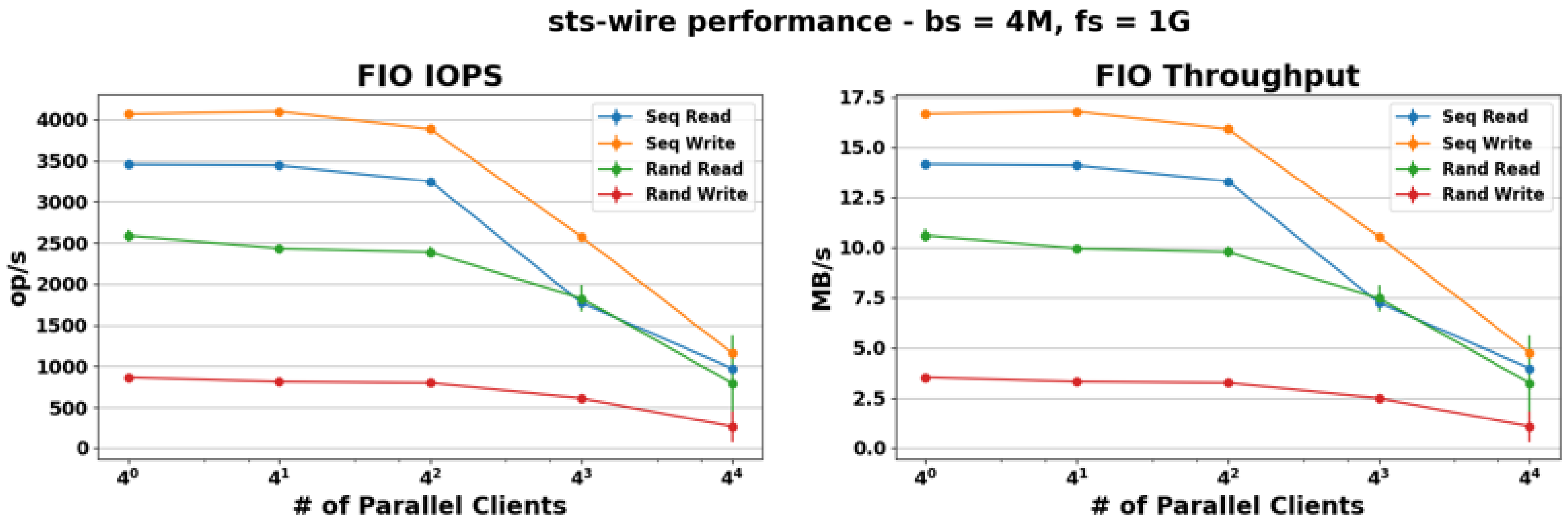
s3fs-fuse performance - bs = 4M, fs = 1G

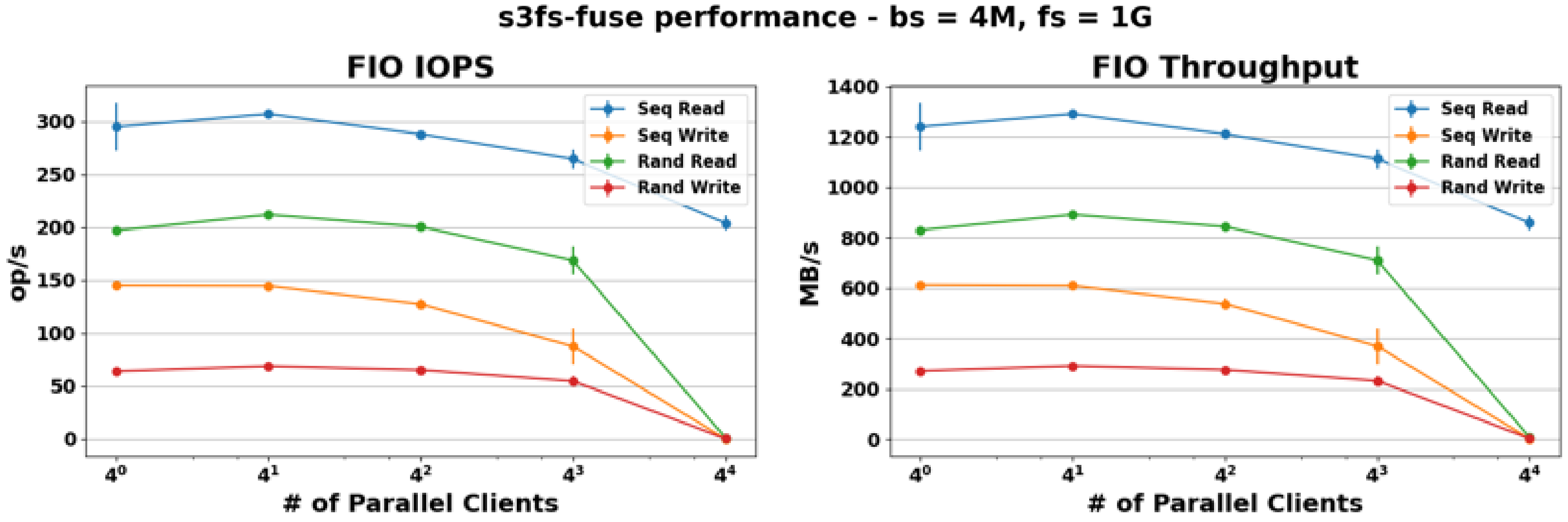


RGW IOPS

RGW Throughput

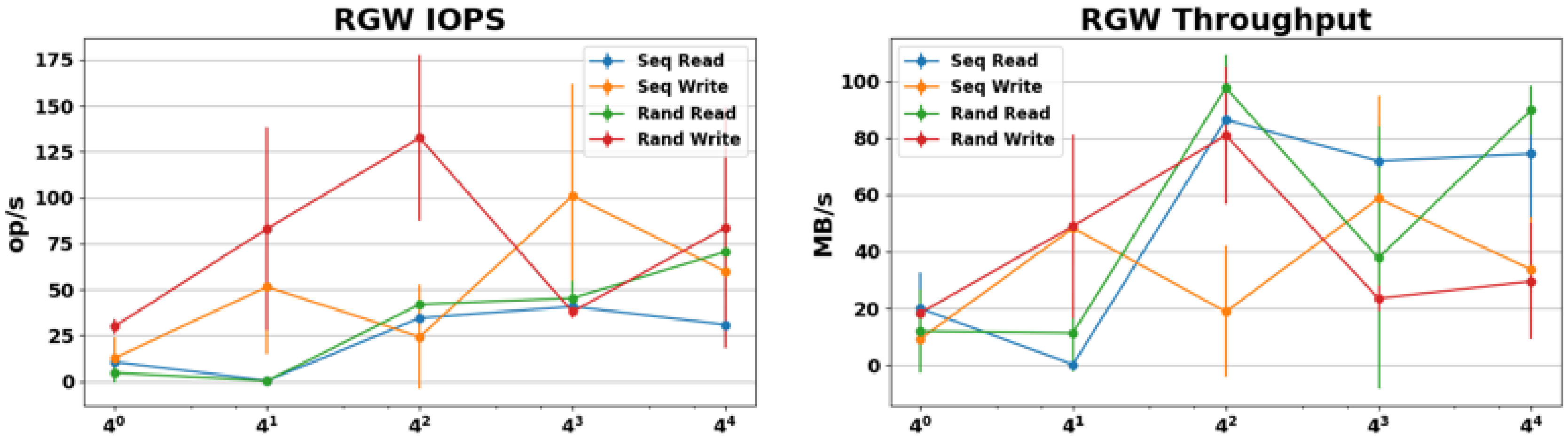
Performance test: blocksize=4 MB



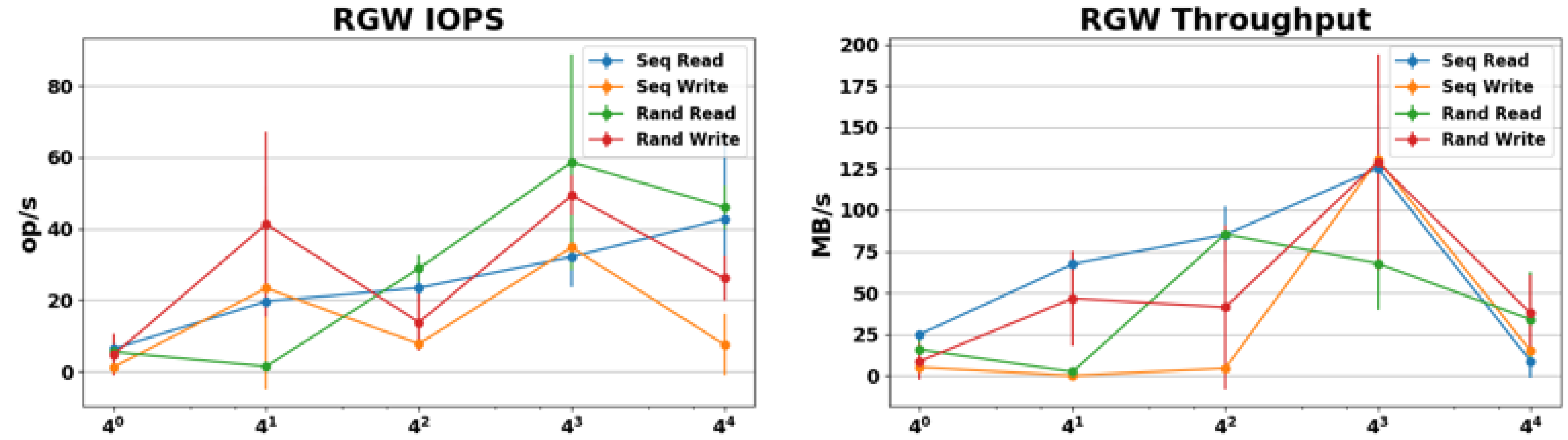


Performance test: blocksize=4 KB

sts-wire performance - bs = 4k, fs = 1G

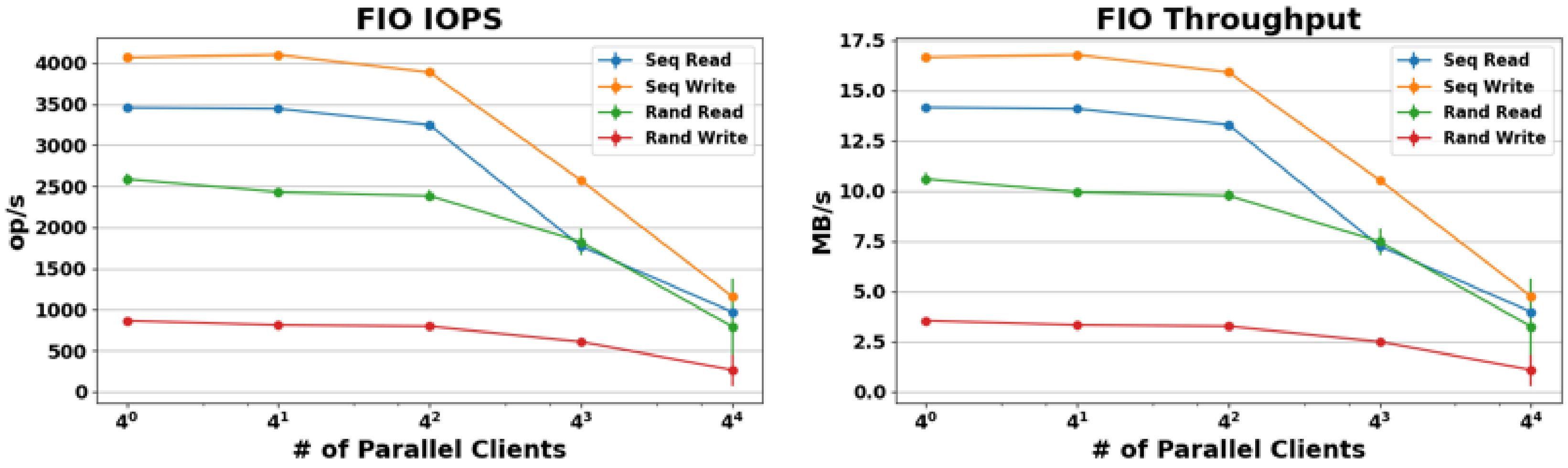


s3fs-fuse performance - bs = 4k, fs = 1G

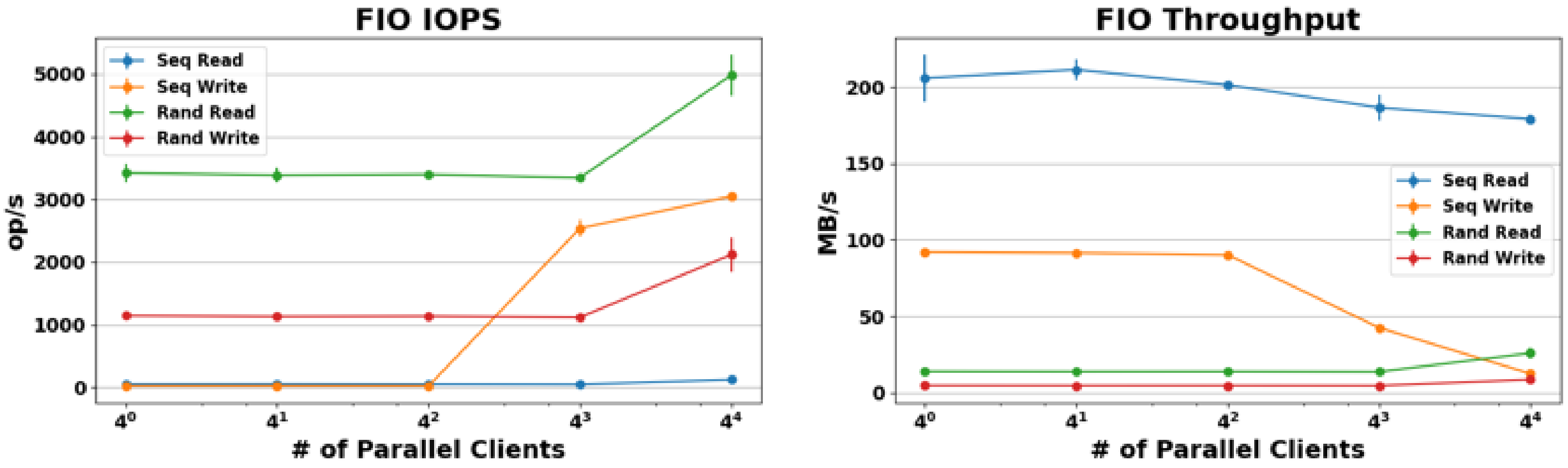


Performance test: blocksize=4 KB

sts-wire performance - bs = 4k, fs = 1G

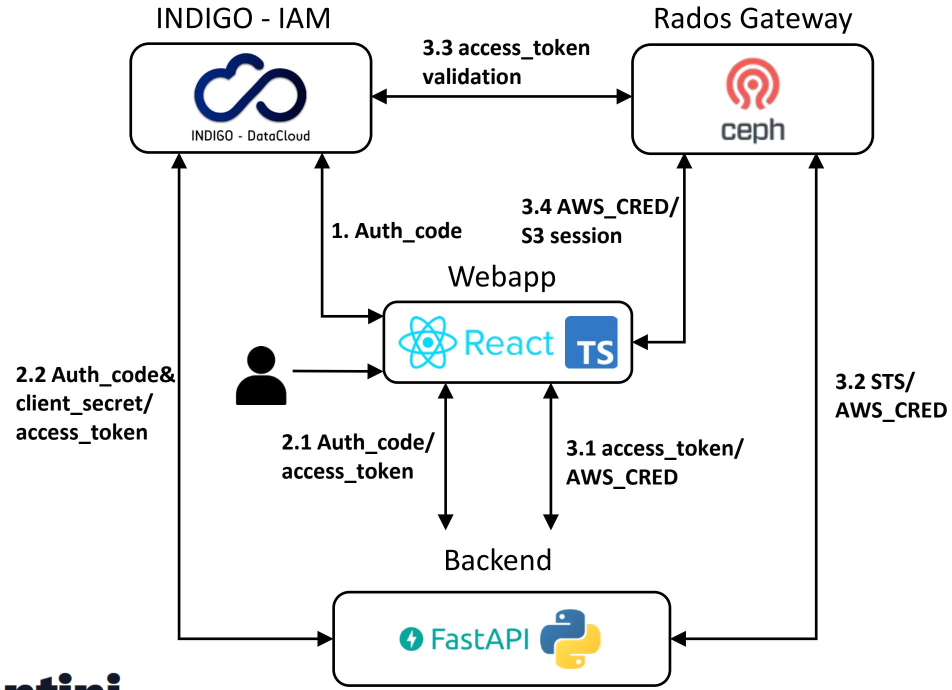


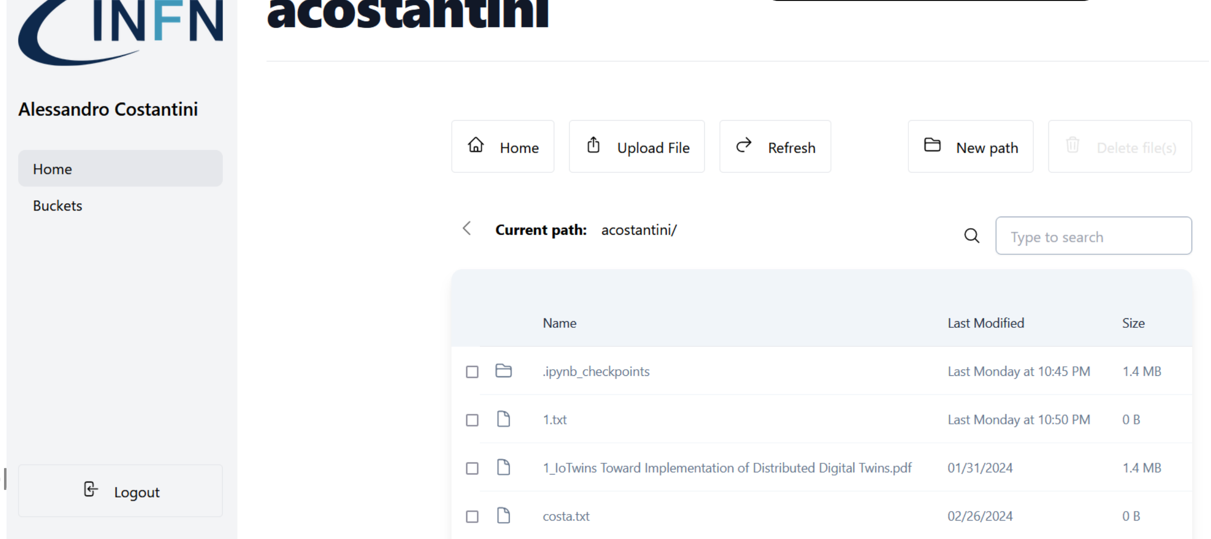
s3fs-fuse performance - bs = 4k, fs = 1G



S3 Web App

- Based on **React and FastAPI**.
- Performs **AuthN/AuthZ** with IAM.
- Uses IAM Access Token to perform **STS with RGW**.
- **S3 operations** using AWS SDK library.




acostantini

Alessandro Costantini

Current path: acostantini/

	Name	Last Modified	Size
□	.ipynb_checkpoints	Last Monday at 10:45 PM	1.4 MB
□	1.txt	Last Monday at 10:50 PM	0 B
□	1_IoTwins Toward Implementation of Distributed Digital Twins.pdf	01/31/2024	1.4 MB
□	costa.txt	02/26/2024	0 B

Conclusion and Future Plans

- Ceph Object storage (**RGW**) deployed using MULTI-SITE configuration
  - To replace the actual MinIO Gateway implementation
  - Gepgraphycal disctribution using Active-Passive approach
- **IAM-CEPH-OPA** integration tested together with the support to STS and Rclone to allow POSIX-like mount of object storage
  - **OPA** offers the possibility to create **highly selective policies**
- A Web app acts as a **GUI** to interact with RGW

- Under implementation
  - Event driven approach using the RGW S3 notification
  - Automate the software distribution with CVM-FS starting from buckets