Istituto Nazionale di Fisica Nucleare
Centro Nazionale per la Ricerca e lo Sviluppo nelle Tecnologie Informatiche e Telematiche

POSIX access to remote storage with OIDC AuthN/AuthZ

Carmelo Pellegrino (carmelo.pellegrino@cnaf.infn.it)

Davide Salomoni (davide.salomoni@cnaf.infn.it)

Federico Fornari (federico.fornari@cnaf.infn.it)

Ahmad Alkhansa (ahmad.alkhansa@cnaf.infn.it)

Alessandro Costantini (alessandro.costantini@cnaf.infn.it)

Introduction

- Several **emerging use cases** of experiments/collaborations needing **local POSIX access** to storage provided by INFN-CNAF:
  - Test-stand TEX for **Eupraxia** asks for **50 TB/year** to archive data to be stored on disk
    - The collaboration **needs** to access data via **POSIX** in read-write mode **from Frascati (Rome)**
    - The software use a single UNIX user reading and writing data on disk
  - **NEWSdm**: "Is it possible to **access** our **storage area** without worrying about token renewal, for example with Rclone?"
  - WLCG experiments would like to **access cloud storage** resources in a **POSIX**-like way
    - Multiple solutions are available (**Ceph, S3, CVMFS, CernBOX**), but which is the most suitable?
  - Many more expected

MinIO with Vault-delegated STS

- **Hashicorp Vault** is a software that allows to **securely store secrets**
- Vault can **interact** with **MinIO** to get temporary **S3 credentials**
  - Unofficial **plugin** by Hashicorp (https://github.com/StatCan/vault-plugin-secrets-minio)
- **Vault** can be configured to be accessed through **OIDC AuthN**
  - Vault **supports Indigo IAM,** the **OpenID Connect provider** developed by INFN-CNAF
- Vault can **supply Secure Token Service (STS)** functionality for MinIO
- A **policy** must be defined in **MinIO** and is **linked** to a **Vault role** to perform operations on **buckets** based on IAM token **groups** claim value

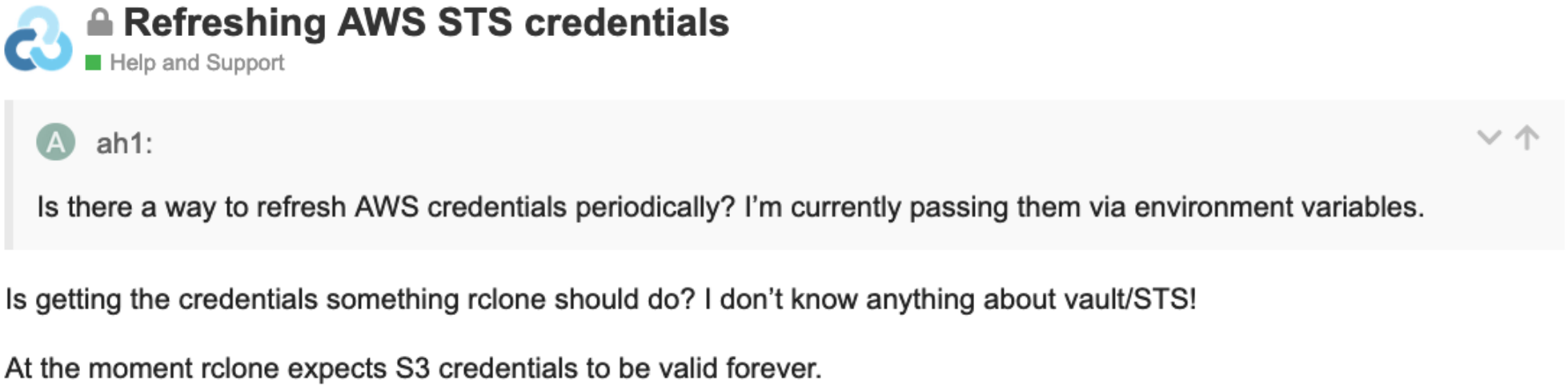
Tested Client Solutions

- **Rclone** 

- **s3fs-fuse** 

Rclone

- **S3 credentials valid for 1h** in this approach, so how to **keep** your locally mounted **bucket connected** to the storage server when **credentials expire**?
- **Client application must** be smart enough to **automatically refresh** temporary **S3 credentials**
- Unfortunately, **Rclone** does **not** fit this requirement (at least for **S3**)



Refreshing AWS STS credentials

Help and Support

ah1:

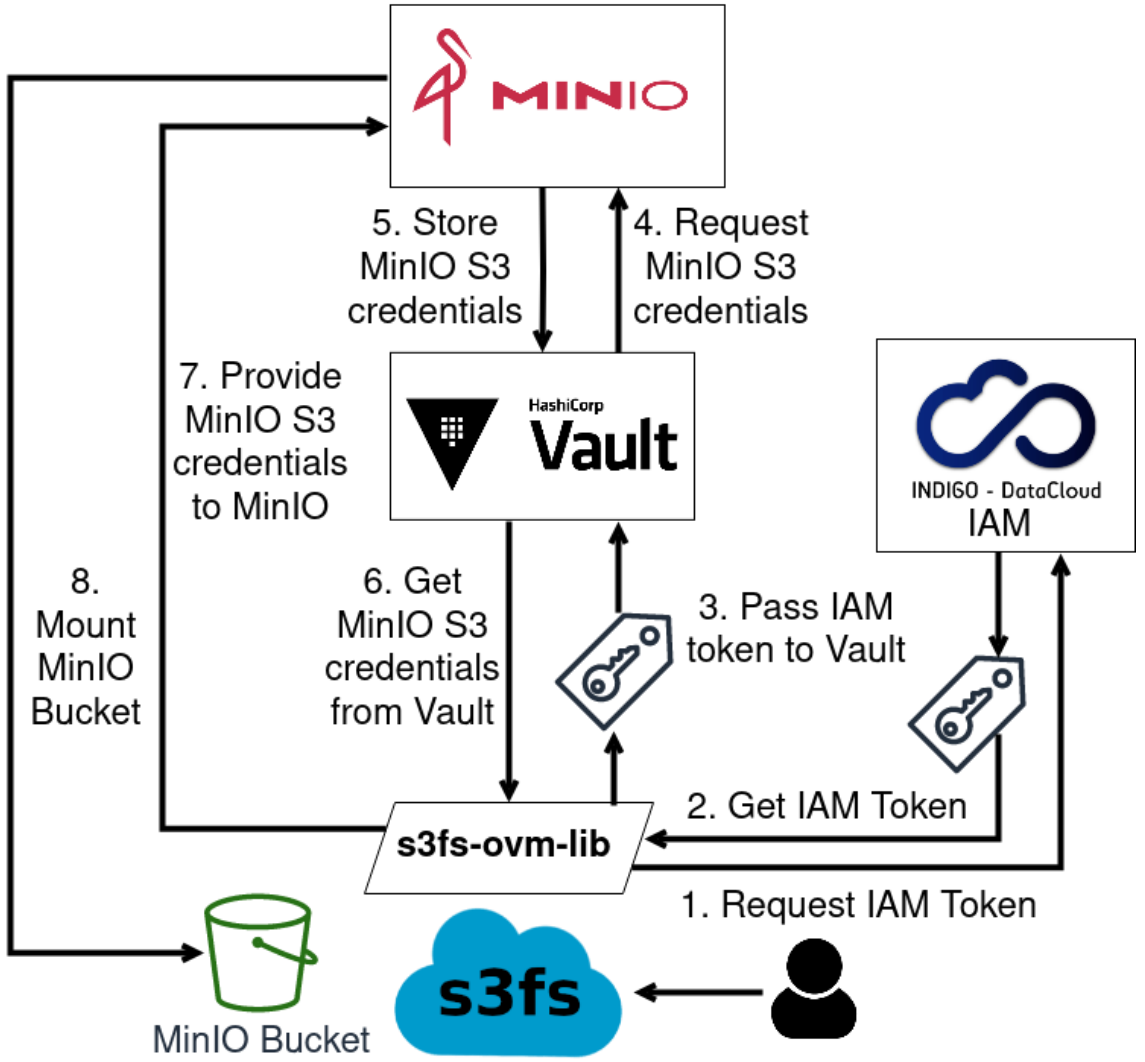
Is there a way to refresh AWS credentials periodically? I'm currently passing them via environment variables.

Is getting the credentials something rclone should do? I don't know anything about vault/STS!

At the moment rclone expects S3 credentials to be valid forever.

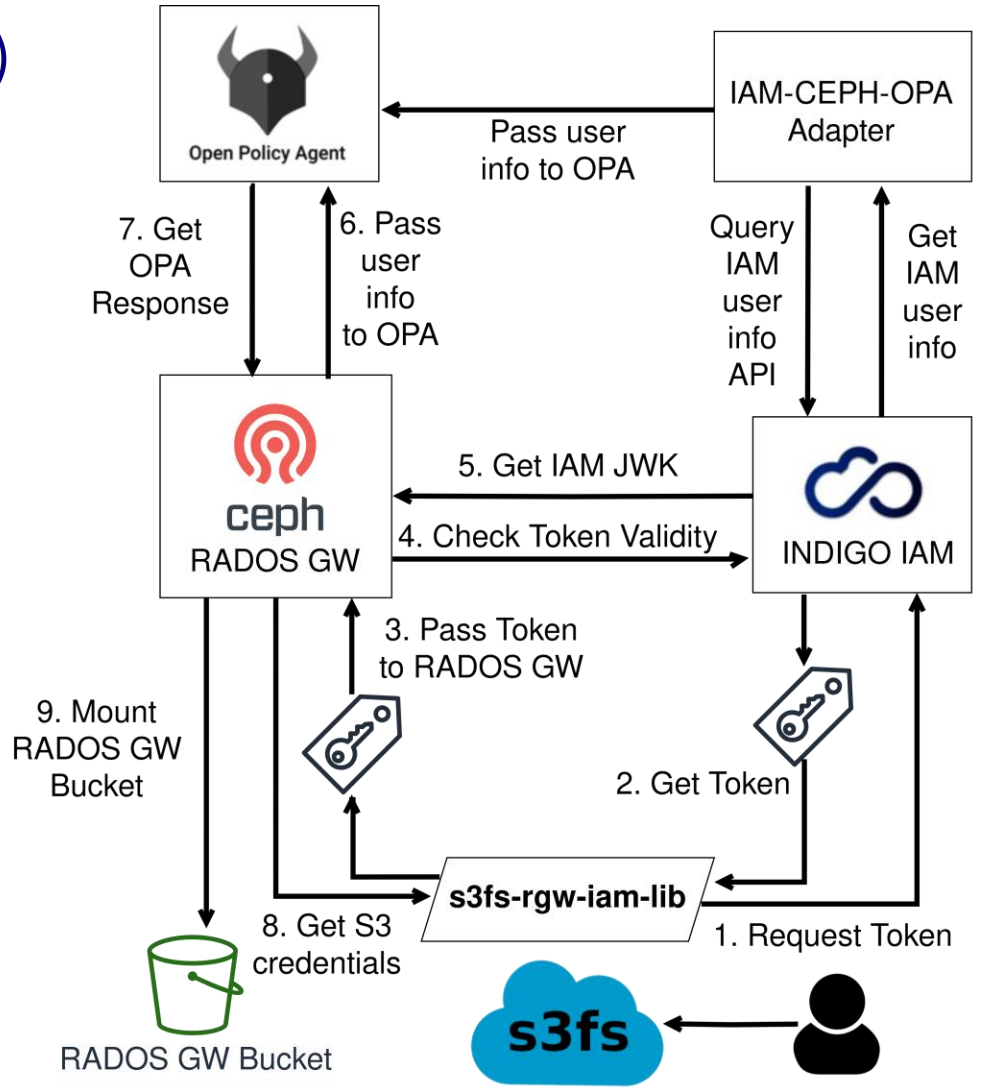
s3fs-fuse (s3fs-ovm-lib)

- s3fs-ovm-lib is a **shared library** (developed in C++) that performs **credential processing** of **s3fs-fuse** using:
  - **oidc-agent** C++ API to get an **access token** from Indigo **IAM**
  - **Vault** C++ API to obtain **S3 temporary credentials** from **MinIO**
  - https://baltig.infn.it/fornari/s3fs-oidc-vault-minio-lib
- s3fs-ovm-lib takes care of **temporary S3 credentials updating** whenever s3fs-fuse detects expiration



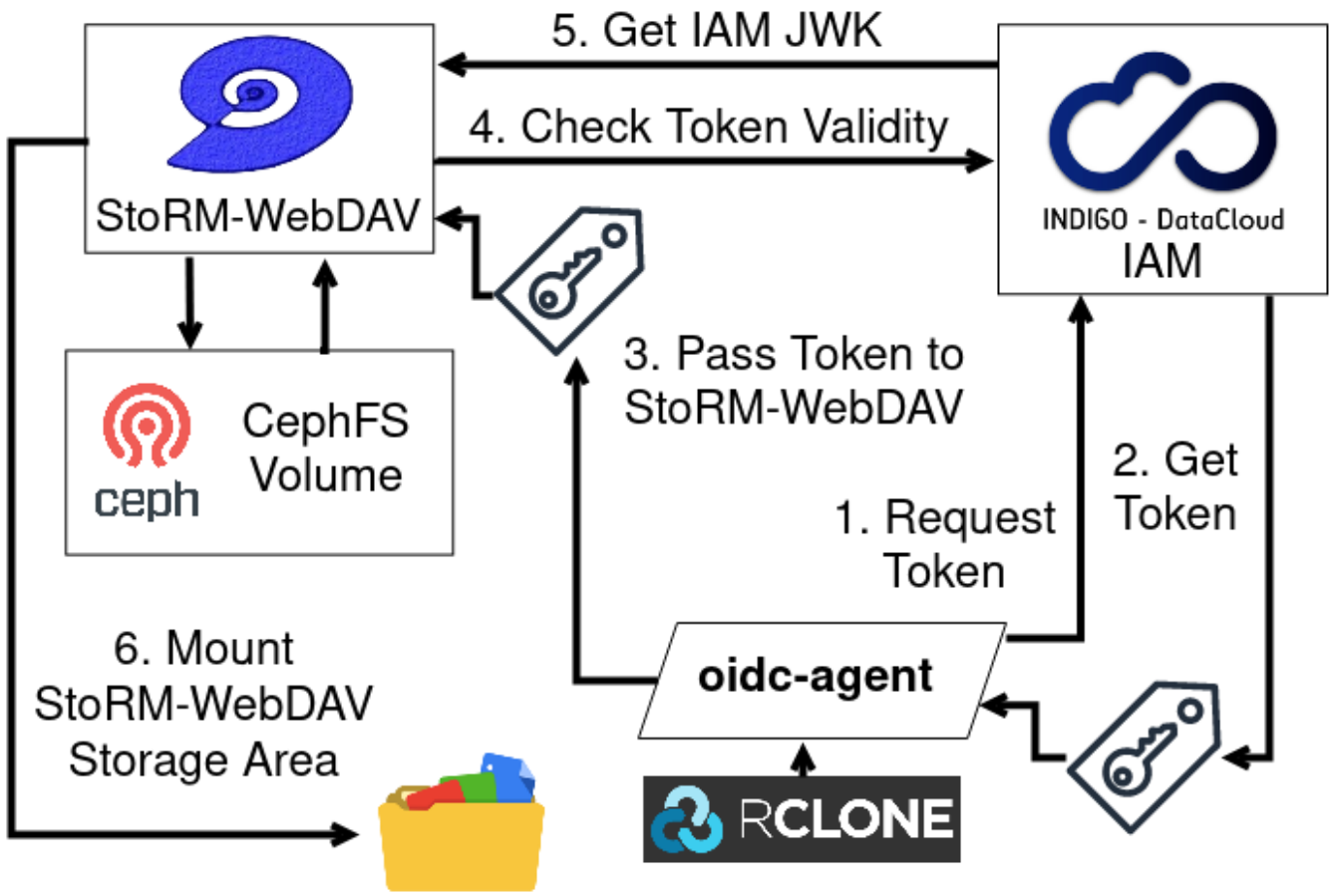
AuthN/AuthZ workflow with s3fs-rgw-iam-lib

- An **additional** C++ **credlib plugin** (s3fs-rgw-iam-lib) has been developed for **IAM** AuthN with **RADOS Gateway**
  - https://baltig.infn.it/fornari/s3fs-rgw-iam-lib
- This library retrieves an **IAM access token** and gives it to Ceph **RGW** requesting for an **S3 operation**
- RGW verifies the **validity** of the **IAM token** and sends the operation request to **Open Policy Agent** in addition to information about the user
- A **IAM-CEPH-OPA Adapter** Python application keeps **OPA updated** with newly created **users information** from IAM
- OPA's response **depends** on the available **policies**
- Upon OPA's affermative response, **s3fs** gets **temporary S3 credentials** and mounts the **bucket**

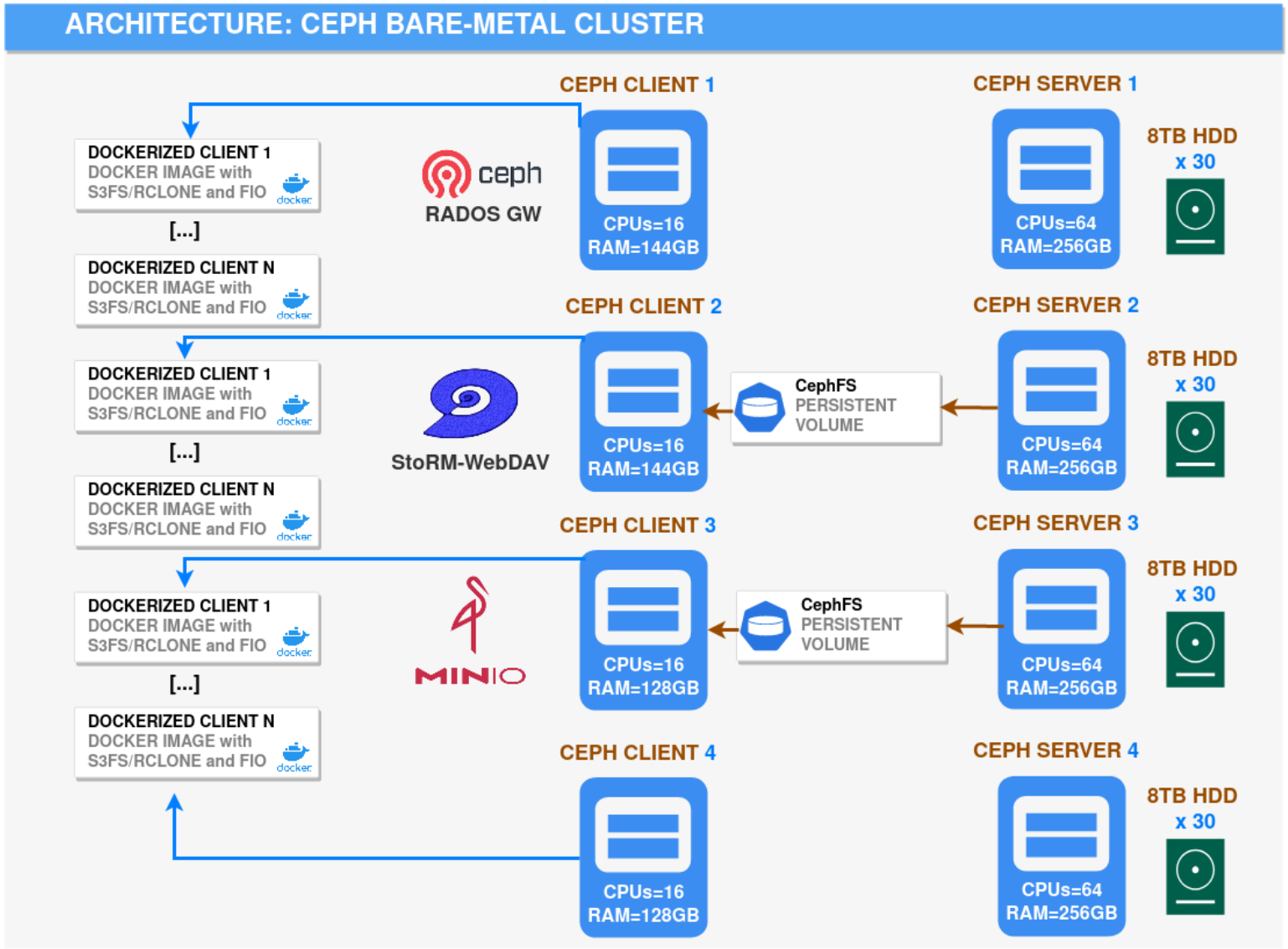


Rclone + StoRM-WebDAV + CephFS

- INFN-CNAF is a **HTTP WebDAV** site (for **non-POSIX** storage)
- **Rclone** can **mount** a **StoRM-WebDAV** storage area (**SA**) providing **POSIX access**
  - For **WebDAV** remote storage, **Rclone allows** the user to **provide a command** (oidc-agent) for the application to **automatically renew tokens**
- **StoRM-WebDAV exports** data from **POSIX** file system (**CephFS**), **no object storage**

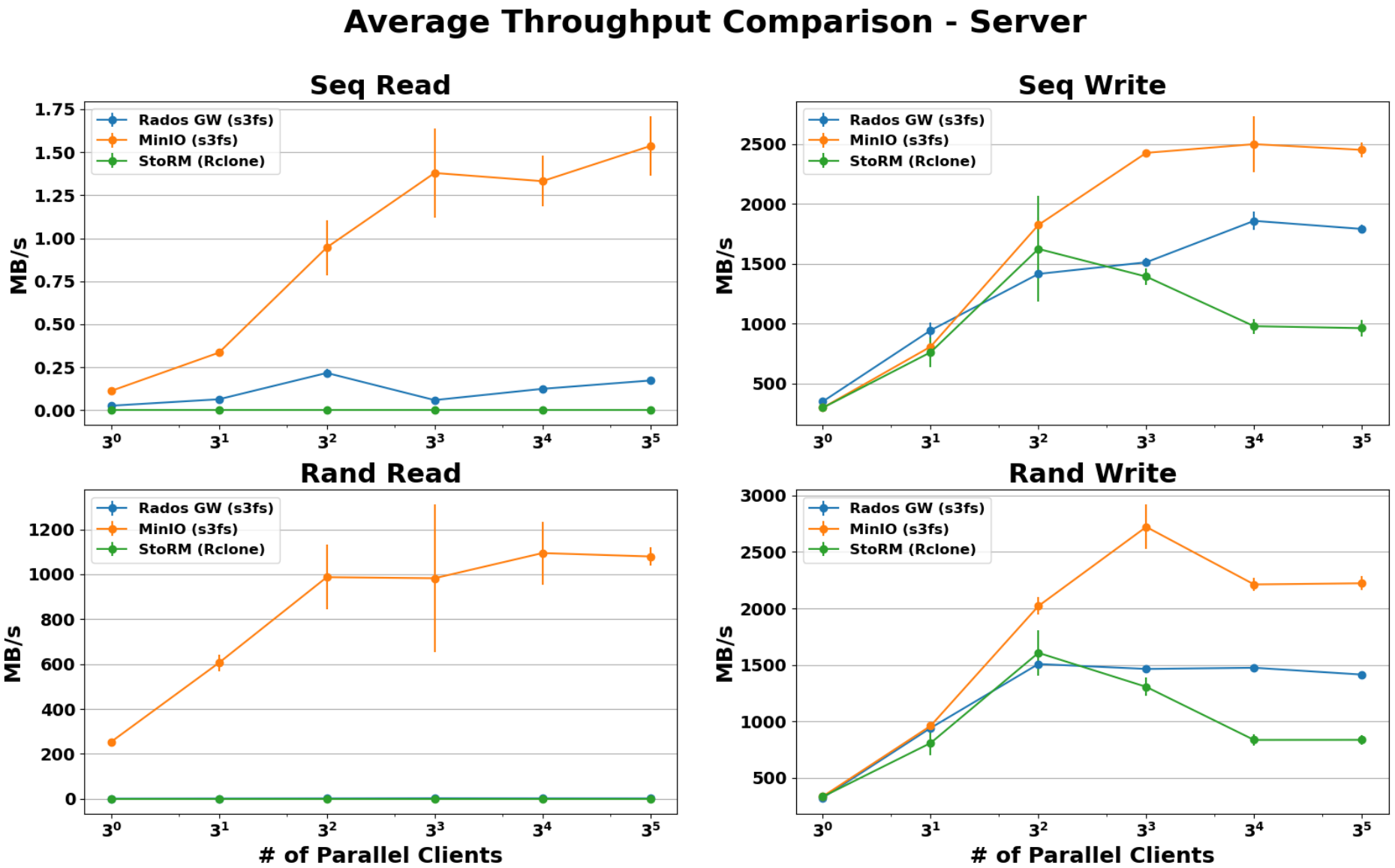


Scalability Tests – Testbed Setup



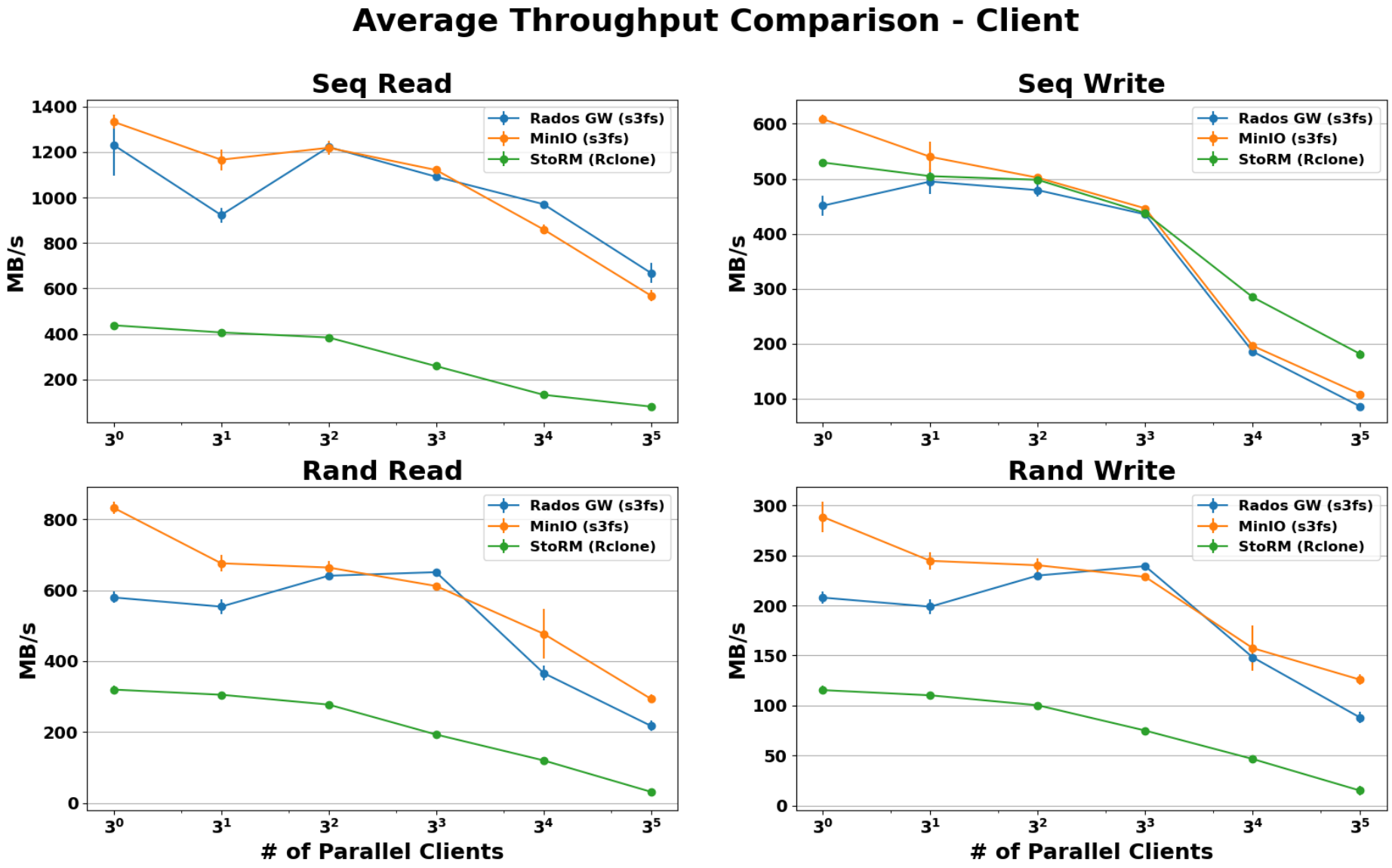
- Ceph testbed:
  - **4 server** nodes
  - **4 client** nodes
  - **2x10 Gbit** NIC per node
  - **120 8TB HDD**
- **3** Ceph client nodes host **gateway services:**
  - Rados GW
  - MinIO
  - StoRM-WebDAV
- **4** Ceph client nodes host client **containers** with **s3fs/Rclone** to mount personal buckets/storage areas and with **fio** to perform tests

Scalability Tests – Server Side Results



- Each **point** in the plots consists of the **mean** and relative **error** of **5 runs**
- Each **run** is a **fio** sequential/random write/read of a single **O(GB)** file **per client**
- **Throughput** seen by Ceph cluster during the tests for the interested **Ceph pool**

Scalability Tests – Client Side Results



- **Throughput** seen by **fio** during the same tests
- **s3fs** (cache-enabled) yields **better read** performance w.r.t. Rclone
- **MinIO + CephFS** generally shows **better throughput** than **RADOS GW**
- **Rclone + StoRM-WebDAV** shows **poorer** results w.r.t. s3fs-fuse **except** for **sequential write**

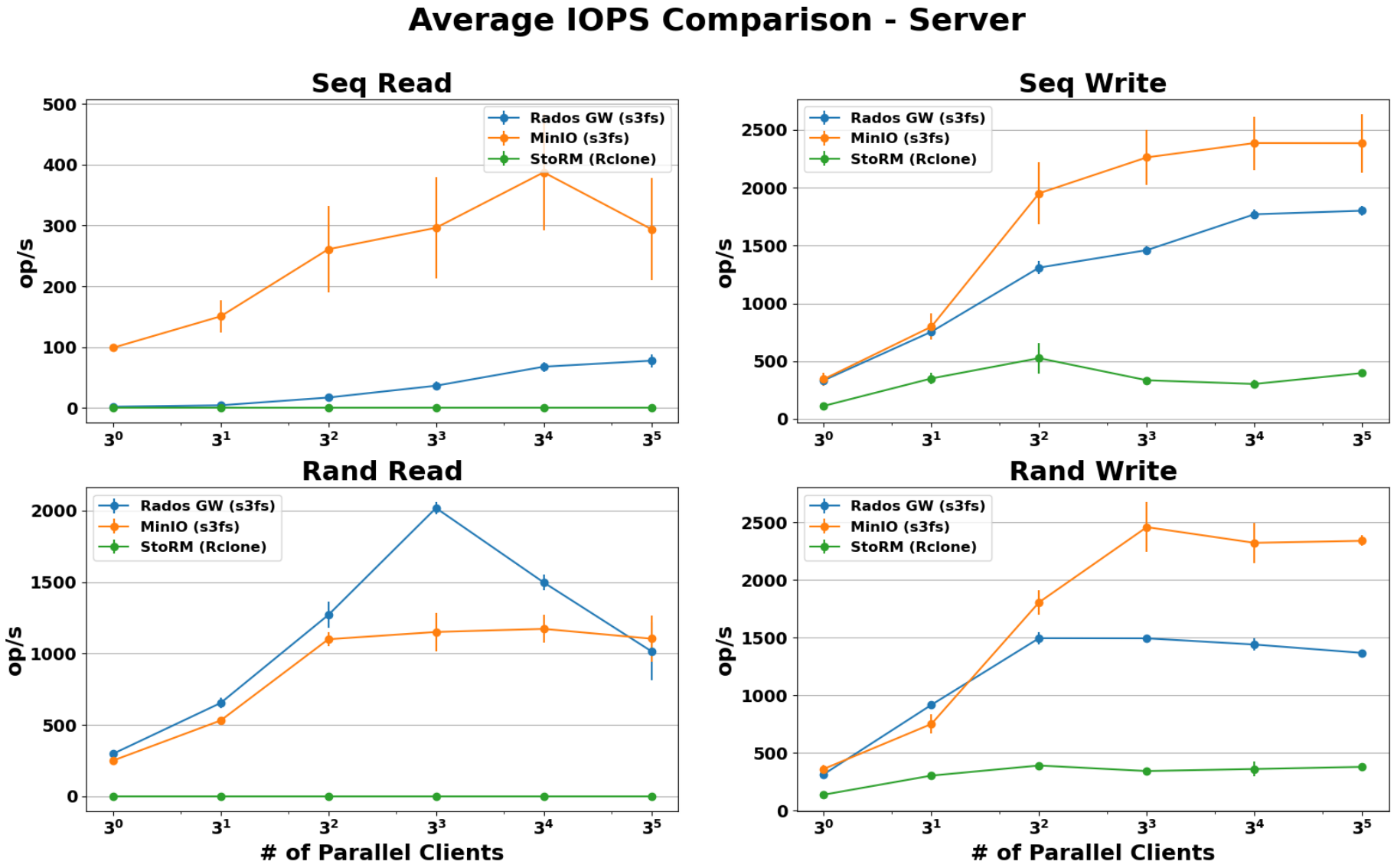
Conclusions and future plans

- **s3fs-fuse** seems to be a **promising** application to **support** the **remote** storage local **mount** with OpenID Connect AuthN/AuthZ mechanism
- **Rclone** can be tuned with a series of parameters, but shows **poor** performance **out of the box** with respect to s3fs-fuse
- **MinIO** in combination with CephFS generally supports **slightly higher throughput** than **RADOS GW**
- **Future** tests may be done **increasing** the number of **client nodes** and involving **alternative WebDAV** storage services for **Rclone** (e.g. **ownCloud**)

THANK YOU VERY MUCH!

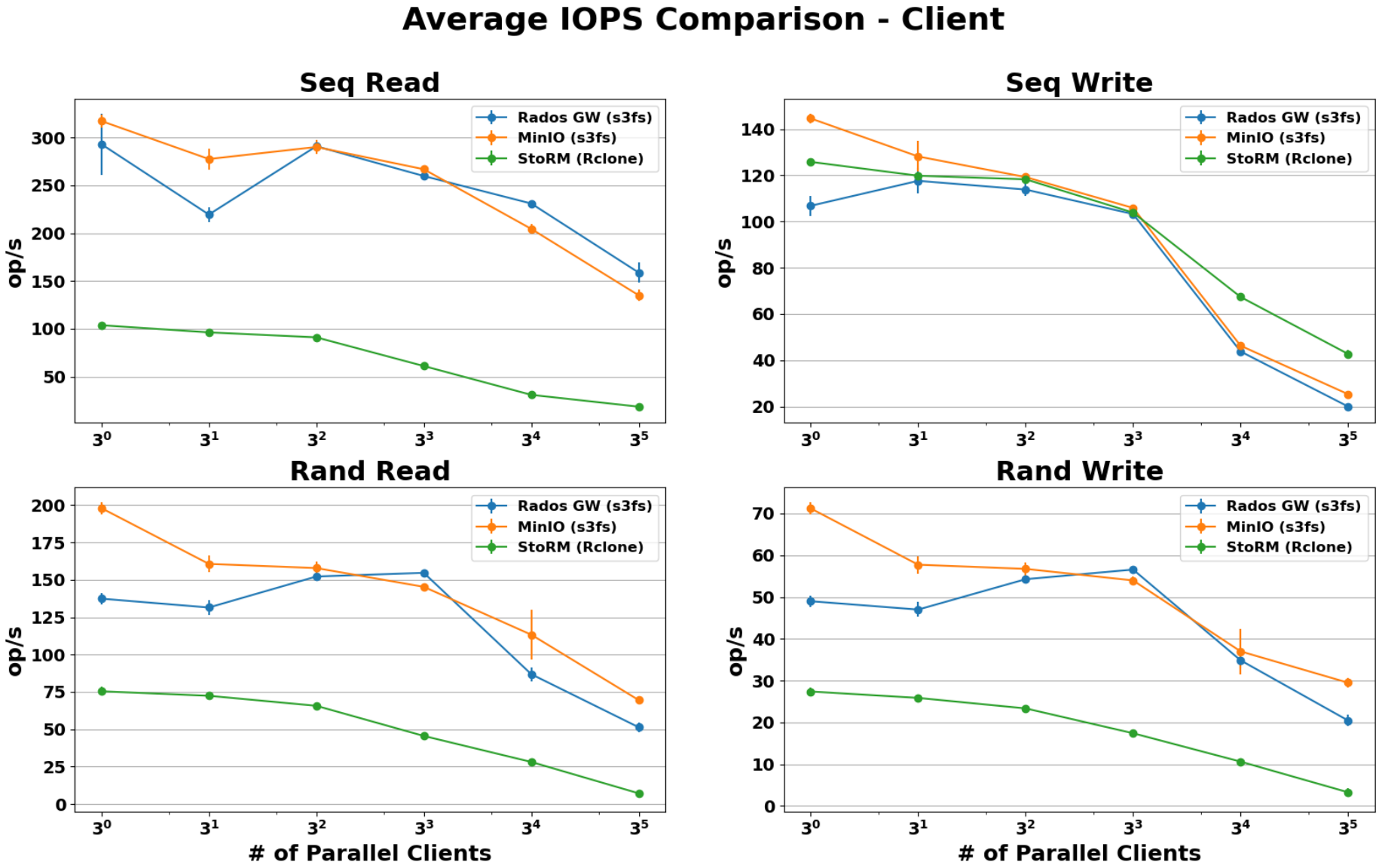
BACKUP SLIDES

Scalability Tests – Server Results



- Each **point** in the plots consists of **mean** and relative **error** of **5 runs**
- Each **run** is a **fio** sequential/random write/read of a single **O(GB)** file **per client**
- These are the **IOPS** seen by Ceph cluster during the tests for the interested **Ceph pool**

Scalability Tests – Client Results



- Each **point** in the plots consists of **mean** and relative **error** of **5 runs**
- Each **run** is a **fio** sequential/random write/read of a single **O(GB)** file **per client**
- These are the **IOPS** seen by **fio** during the performance tests