# Declarative Object Storage at Scale

Integrating Rook, Ceph, and OpenStack

Joachim Kraftmayer, Artem Torubarov FOSDEM Feb, 2025

# NeoNephos | TINUX | Europe



**Foundation in formation** 

















aufgrund eines Beschlusses des Deutschen Bundestages







#### Blueprint for Open Cloud

#### **Free and Open**

All parts and components are OSS. Actively maintained.



#### Blueprint for Open Cloud

#### **Free and Open**

All parts and components are OSS. Actively maintained.

#### **Seamlessly integrated**

Tested together. Reproducible. Declarative configuration.



#### Blueprint for Open Cloud

#### **Free and Open**

All parts and components are OSS. Actively maintained.

#### **Seamlessly integrated**

Tested together. Reproducible. Declarative configuration.

#### Maintained

Operated in production with the latest releases.



#### Case study

100

**Deployments** 

16

Regions

>5Pb

Cluster size

1-2

Availability Zones

>100

Nodes per cluster

>1000

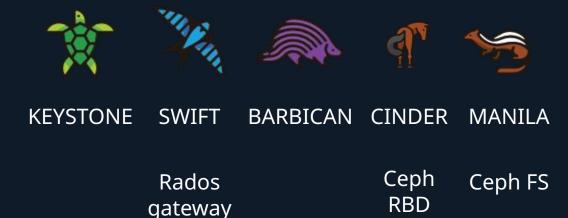
OSDs per cluster























(n) ceph

KEYSTONE

SWIFT

BARBICAN

CINDER M.

MANILA

Rados gateway Ceph RBD Ceph FS



ROOK

Object Storage

Ceph CSI







Rados gateway



Object Storage













KEYSTONE

SWIFT S3 BARBICAN

Rados gateway

Object Storage

#### Part 1: Ceph and Openstack

- Openstack SWIFT is eventually-consistent unlike Ceph
- Keystone V3 API was added in Ceph Jewel (2016)
- RGW SSE and Barbican in Ceph Luminous (2017)
- RGW SWIFT implementation is less mature than S3 today
- RGW is capable of replacing openstack SWIFT

- Public buckets with SWIFT & Keystone [Fixed in reef]
- RGW and Openstack SWIFT ACLs are different [TBD]
- Missing last\_modified field for SWIFT container [WIP]
- SWIFT container temp\_url is not supported <a href="https://recommons.org/linearing-nots-uppercase-nots-nots-uppercase-nots-nots-nots-uppercase-nots-nots-uppercase-nots-nots-uppercase-nots-nots-uppercase-
- Different LastModified format in S3 HeadObject and
  - ListObjectVersions <a>[WIP]</a>

- SWIFT role-based ACL is not supported [TBD]
- S3 Conditional writes <a>[TBD]</a>
- Keystone request caching issue [WIP]
- SHA256 Support in SWIFT temp\_url [fixed in Quincy]
- Container GET and HEAD return different obj count [TBD]
- Cache for Barbican keys [WIP]

- Access S3 objects with SWIFT api if SSE enabled [TBD]
- Versioned S3 bucket incompatible with versioned SWIFT container <a href="[TBD]">[TBD]</a>
- Quota and Usage is tracked only per project and per bucket
- SWIFT Delete account method [TBD]

# Part 2: Declarative Ceph deployment with Rook

Rook allowed us to have fully declarative Object Storage setup:

- 1. Label k8s node with topology labels
- 2. Declare device filters for Rook OSD discovery
- 3. Declare OSD pools
- 4. Declare RGW instances

- Support RGW target placements <u>[done in v1.15]</u>
- Multi-instance RGW deployments <u>[done in v1.16]</u>
- Custom hostname topology label <u>[WIP]</u>
- Support external arbiter mon [WIP]
- Observability: Metrics and tracing. [Improved in Ceph Squid]

#### Object Storage Roadmap

- Maintaining RGW SWIFT and make it in par with RGW S3
- Start discussion about Quota and Usage improvements.
  - Manage quota per user per target placement.
- Test scalability and day-two operations with Rook.
- Data migration.
- Want to hear about your use-cases, looking for case studies

# Tomorrow, **AW1.120 13:35**

**Building Europe's Platform Mesh: Cloud-Native APIs** for Multi-Provider Integration and **Digital Sovereignty** 



