WORKSHOP GARR 2022
NET MAKERS

# Ceph per lo storage Cloud GARR

Alberto Colla
per il Dipartimento CSD GARR

# Table of Contents

- Overview of the GARR Cloud
- What is Ceph
- Why Ceph
- Architecture
- Deployment
- Monitoring
- Experience
- Conclusions

WORKSHOP GARR 2022 NET MAKERS

# GARR Storage and Computing Department (CSD)

Role:

- **provider** of resources ("long tail of science")
- resource **aggregator** (**federation**)
- embody a **replicable** model for storage & computing provisioning

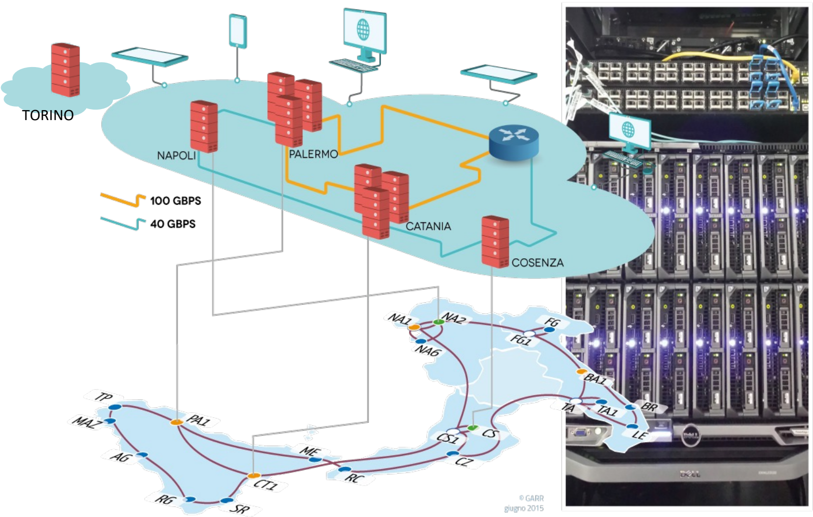
Goals:

- **harmonize** (SSO / federation of resources)
- build **secure** and **open** infrastructures
- enhance **user experience**

# GARR Computing and Storage Infrastructure

## Overall

- CPUs: 7600 cores
- RAM: 60 TB
- **Storage: 13.5 PB** 15% **SSD**
- GPUs: 333 TFLOPS
- Datacentres: 5

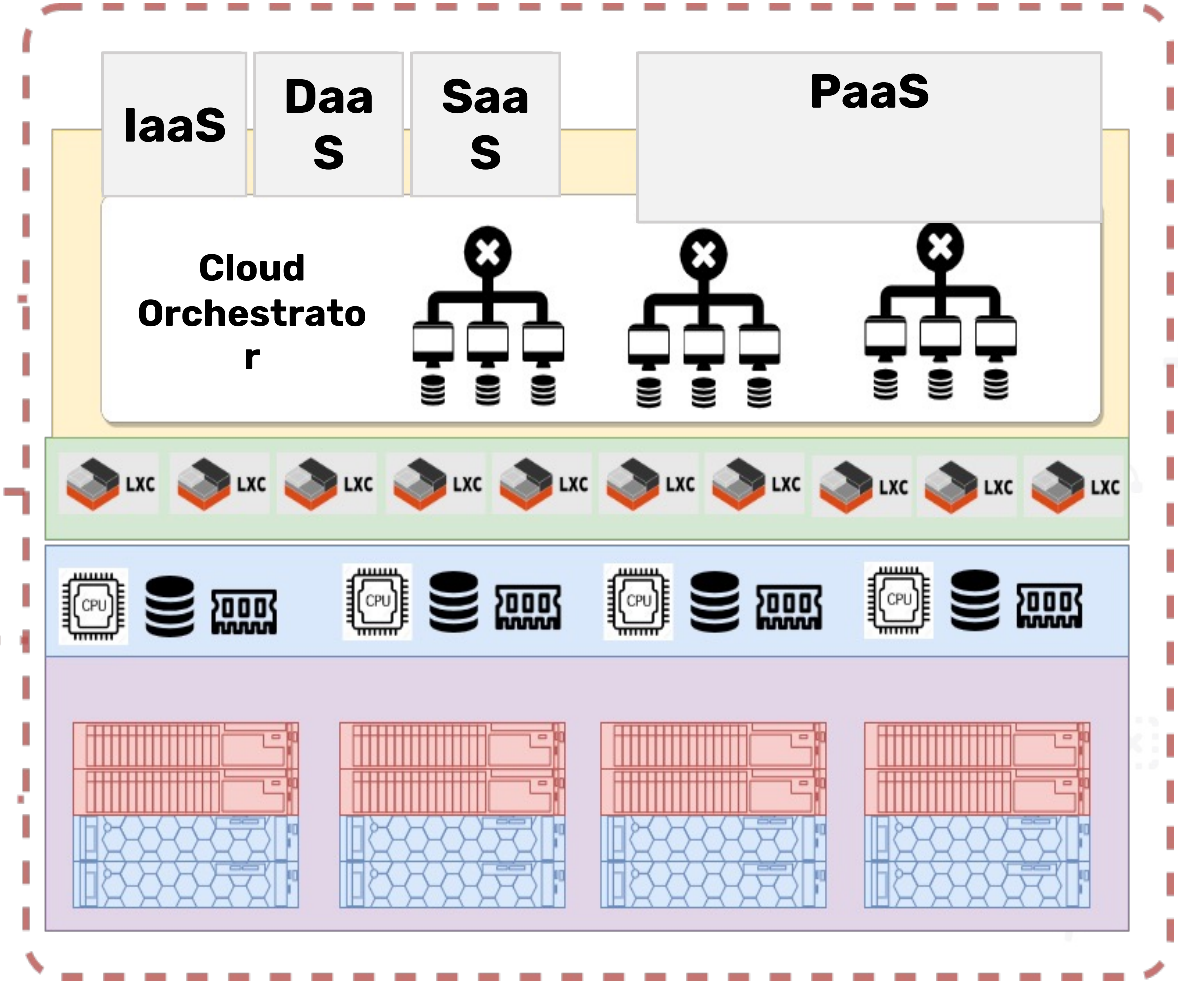


## Storage servers

- **DELL PowerEdge R740XD2**
- 48 HT cores, 384 GB RAM
- 3x 1.9TB SSD Mixed Use ➡️ RocksDB
- **5x 7.68 TB SSD** Read Intensive
- **14x 18TB HDD** 7Krpm

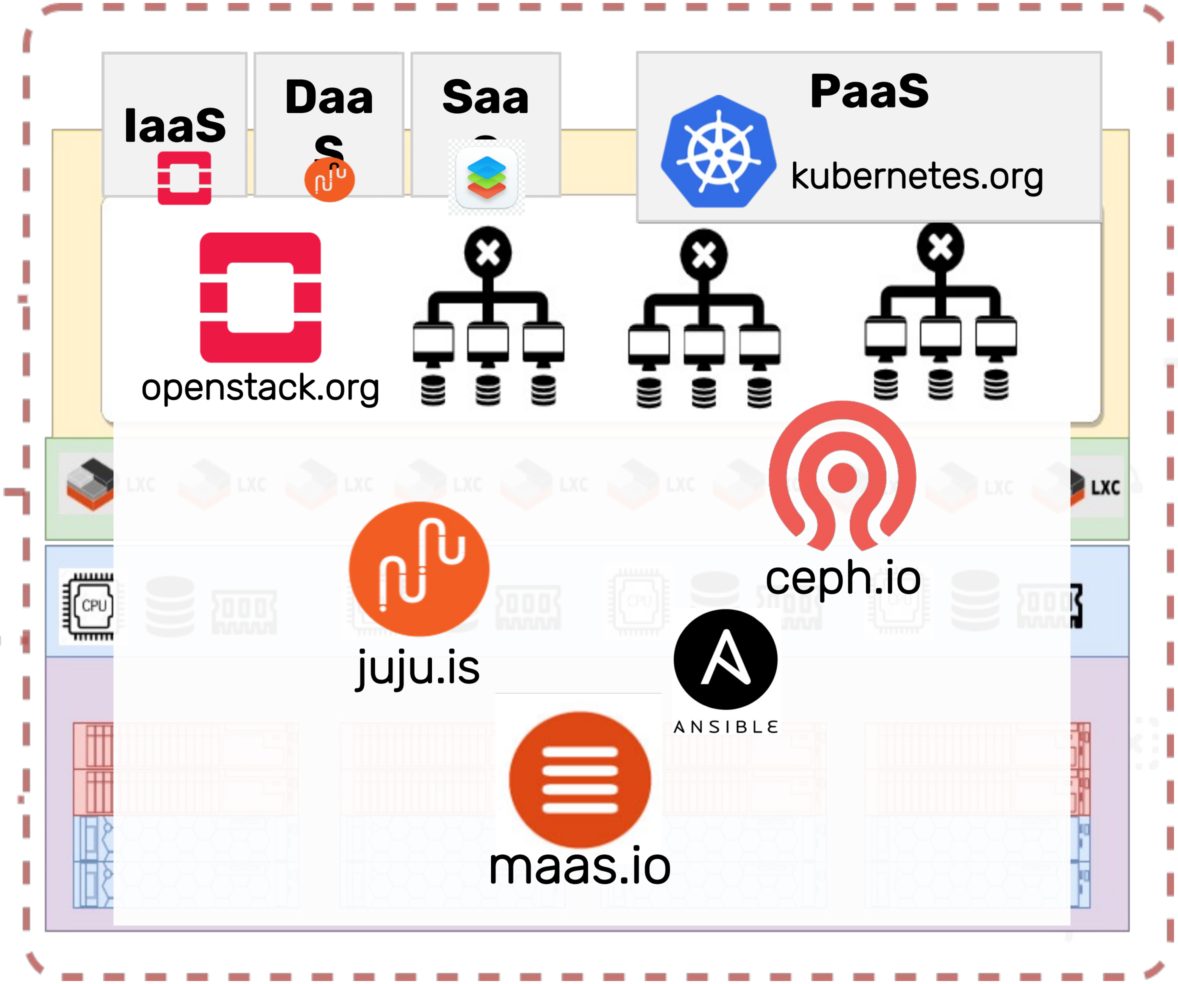
# GARR Cloud Infrastructure: 4-Layers *recipe*

1. **Application Service**
2. **Infrastructure virtualization**
3. **Operating System**
4. **Physical Resources**



# GARR Cloud: the engines

1. **Application Service**
2. **Infrastructure virtualization**
3. **Operating System**
4. **Physical Resources**



# OpenStack

**Horizon** (Dashboard)

**Keystone** (Identity Management)

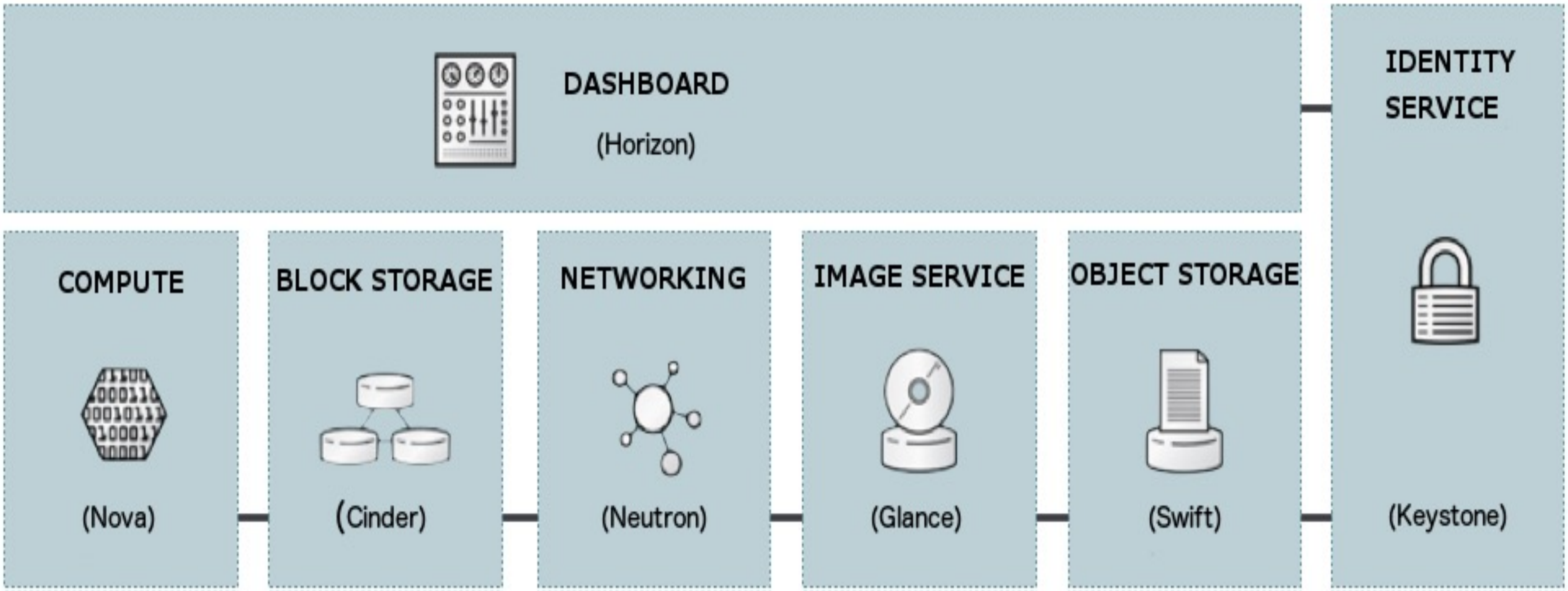
**Nova** (Compute, where VMs are run)

**Glance** (Image Service)

**Cinder** (Block Storage, persistent storage for VMs)

**Swift** (Object Storage, snapshots and not frequently updated data)

**Neutron** (Networking and SDN)

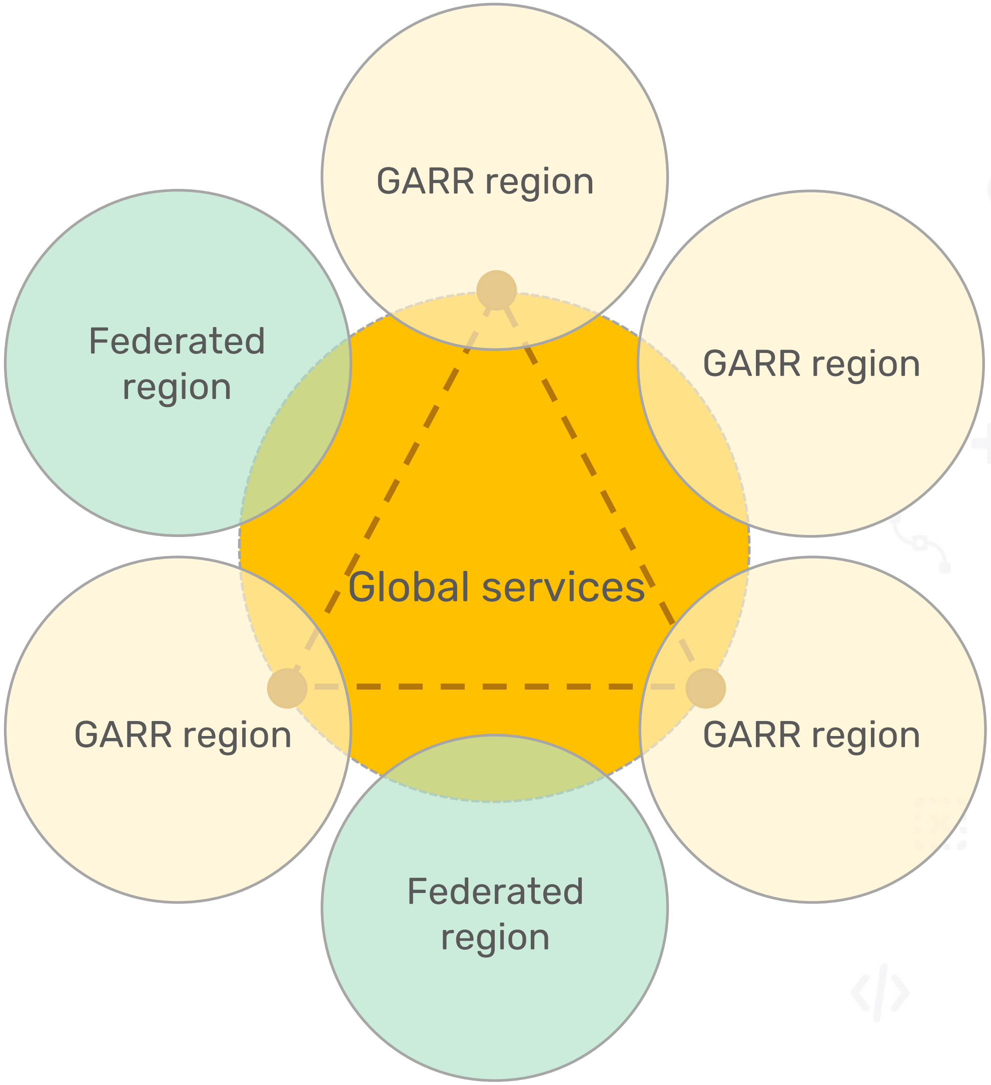


# GARR Cloud architecture

Global services
- ✓ Identity
- ✓ Images
- ✓ Object Store
- ○ geo-distributed
- ○ DNS HA

GARR regions
- ✓ Compute
- ✓ Network
- ✓ Block Store
- ○ Share Identity, Images, Object Store
- ○ Local services

Federated regions
- ✓ Compute
- ✓ Network
- ✓ Block Store
- ○ Share Identity
- ○ Local services
- ○ Managed by You



# OpenStack and Ceph

**Horizon** (Dashboard)

**Keystone** (Identity Management)

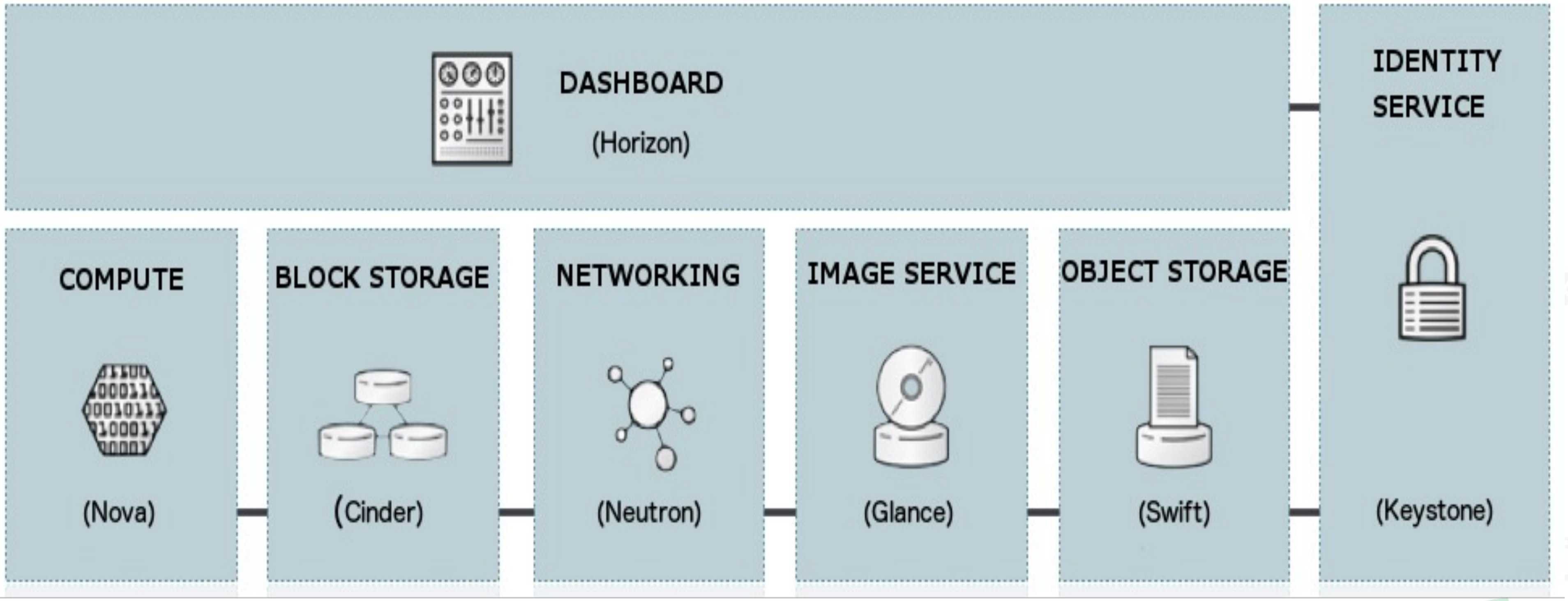
**Nova** (Compute, where VMs are run)

**Glance** (Image Service)

**Cinder** (Block Storage, persistent storage for VMs)

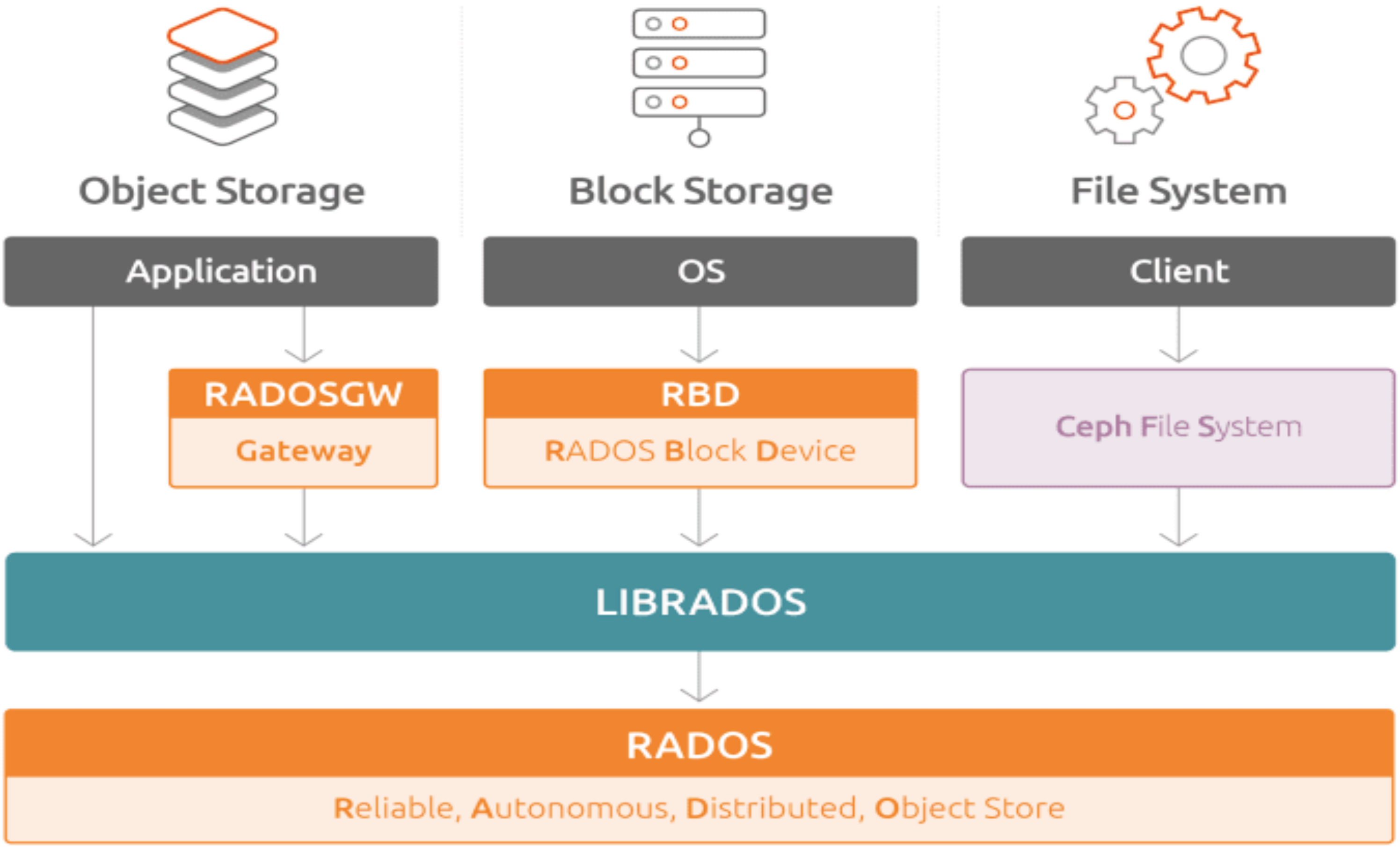
**Swift** (Object Storage, snapshots and not frequently updated data)

**Neutron** (Networking and SDN)

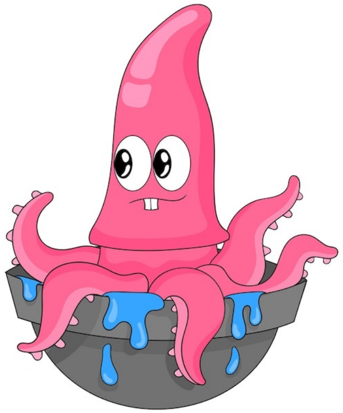


# What is Ceph

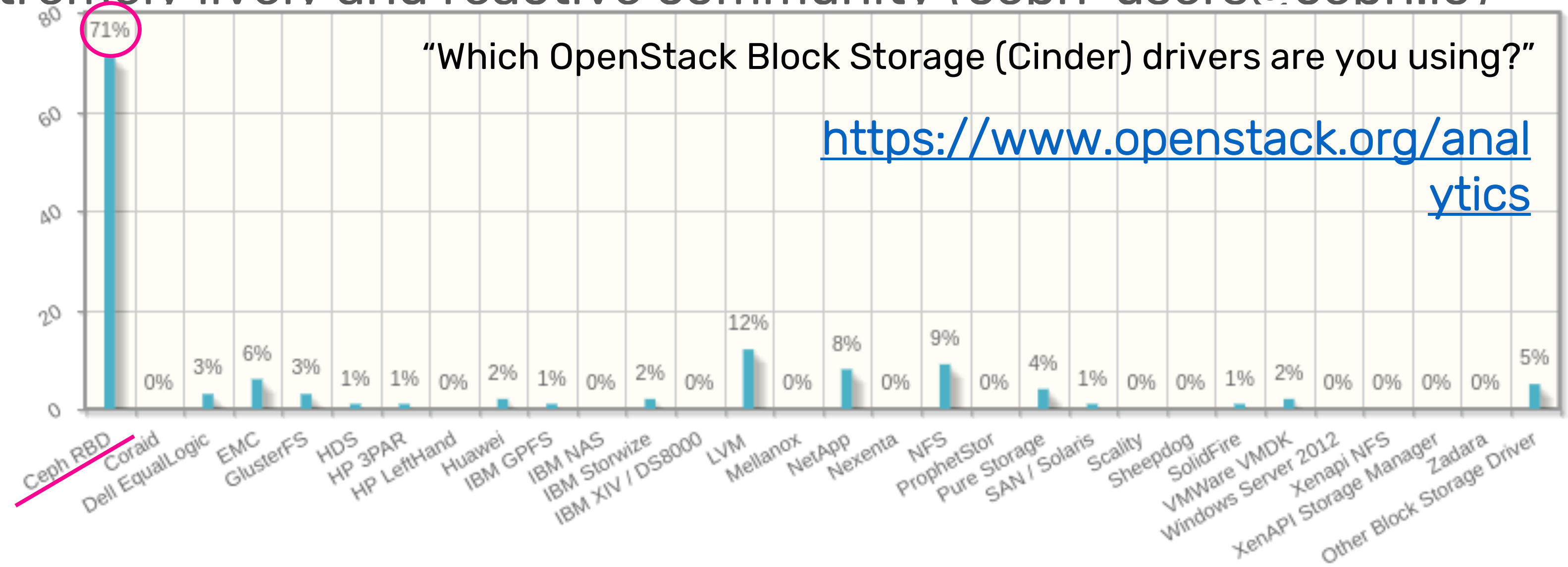
Ceph is **open source software** designed to provide highly scalable **object-**, **block-** and **file-**based storage under a unified system



# Why Ceph



- Open-source, distributed storage
- Lack of SPOF (single point of failure)
- Runs on commodity hardware
- Aggregates any server with any network and disk setup
- Consistently evolving, new functionalities and several improvements
- Extremely lively and reactive community (ceph-users@ceph.io)



"Which OpenStack Block Storage (Cinder) drivers are you using?"

https://www.openstack.org/analytics

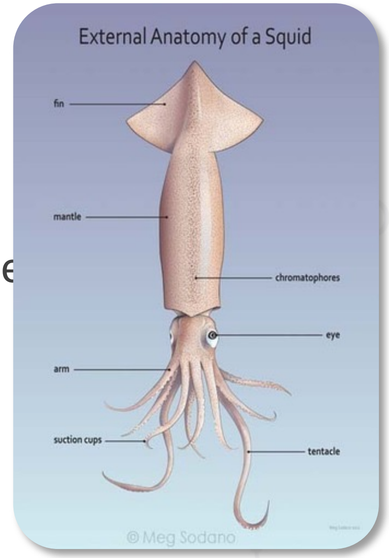
# A little Ceph glossary

- **Object Store Device (OSD)**: the physical disk (plus a slice of CPU/RAM to manage it)
- **Monitors (Mon)**: maintain the map of the cluster state, keeping track of active and failed cluster nodes, cluster configuration, etc.; handle such map to clients
- **Managers (Mgr)**: maintain cluster runtime metrics, enable dashboarding capabilities, provide interface to external monitoring systems
- **Rados Block Device (RBD)**: Ceph’s block storage component
- **Rados Gateway (RGW)**: Ceph’s object storage APIs (swift and S3)
- **Metadata Servers (MDS)**: store metadata for the Ceph File System

- Ceph stores data (objects) within **pools**.
- Pools ensure data redundancy (**Replica** vs. **Erasure-Coding**)
- Within a pool, objects are mapped to **Placement Groups** (**PG**), and placement groups to the OSDs
- OSDs belong to **Device Classes**: default (“hdd”, “ssd”) or custom (“test”, “big”)

# A little Ceph glossary

- The **CRUSH** (Controlled Replication Under Scalable Hashing) algorithm determines how to store and retrieve data
  - The **CRUSH map** describes the physical topology of the cluster: it is passed to clients who can then interact directly with the cluste
  - The **CRUSH rules** define data placement policy, e.g.: what device class to use, how many chunks per rack/host,...



- Two examples of CRUSH rules:

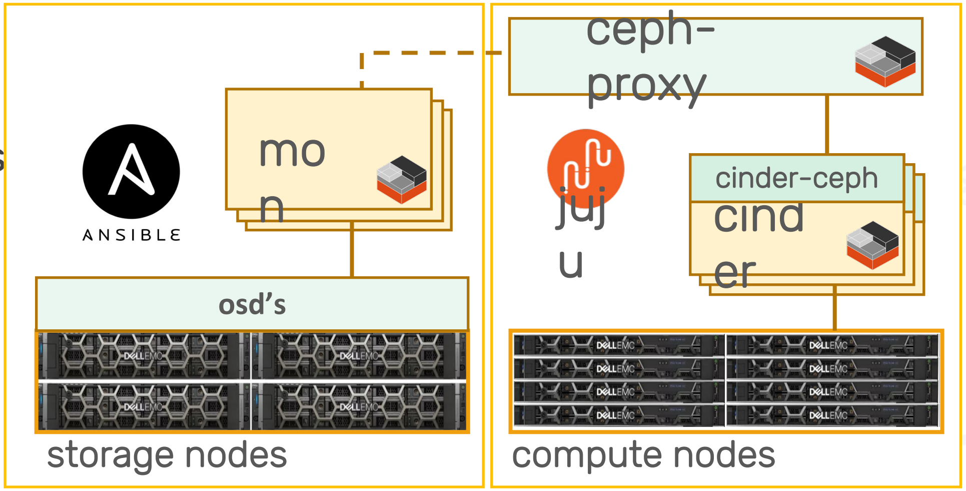
```
rule ssd_rule {
     id 1
     type replicated
     min_size 1
     max_size 3
     step take default class ssd
     step chooseleaf firstn 0 type rack
     step emit
}
```

```
rule default.rgw.buckets.data {
     id 6
     type erasure
     min_size 3
     max_size 10
     step set_chooseleaf_tries 5
     step set_choose_tries 100
     step take default class big
     step choose indep 5 type host
     step chooseleaf indep 2 type osd
     step emit
}
```

# Ceph deployment - two cases

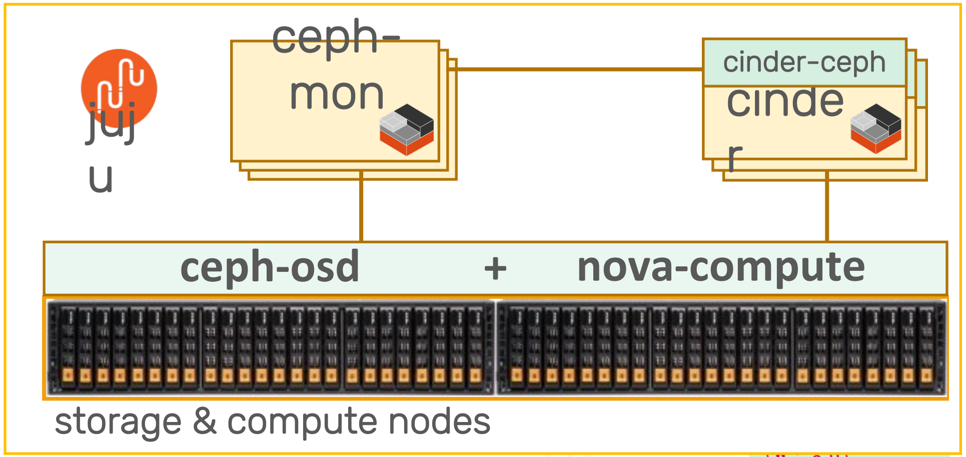
**Catania, Palermo, Napoli regions**

- Ceph w. **ceph-ansible** on storage nodes OpenStack with **Juju** on compute nodes
- Ceph-OpenStack connection via **ceph-proxy** charm
- Ceph-proxy
  - Is configured with **Mon IP addresses** and **ceph-admin credentials**
  - Is related to the Openstack modules
  - Creates users, pools etc. in Ceph according to the directives of the related OpenStack modules
  - passes config parameters to Ceph clients on the modules

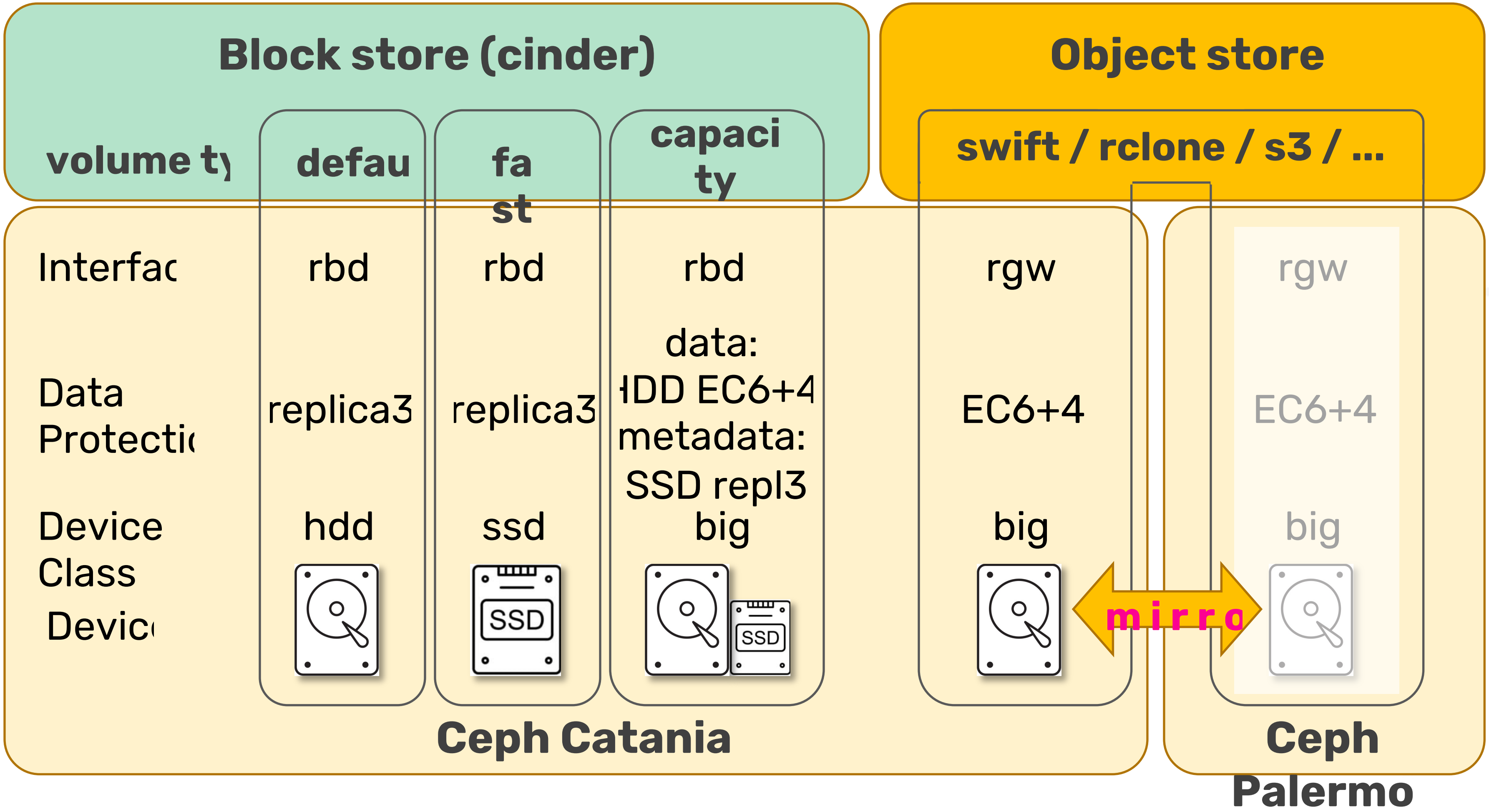


**Torino region:**

- Hyper-converged systems (storage&compute)
  - few servers, but powerful
- Deploy Ceph and Openstack with Juju
- Ceph-mon charm directly connected with OpenStack modules (juju relations)

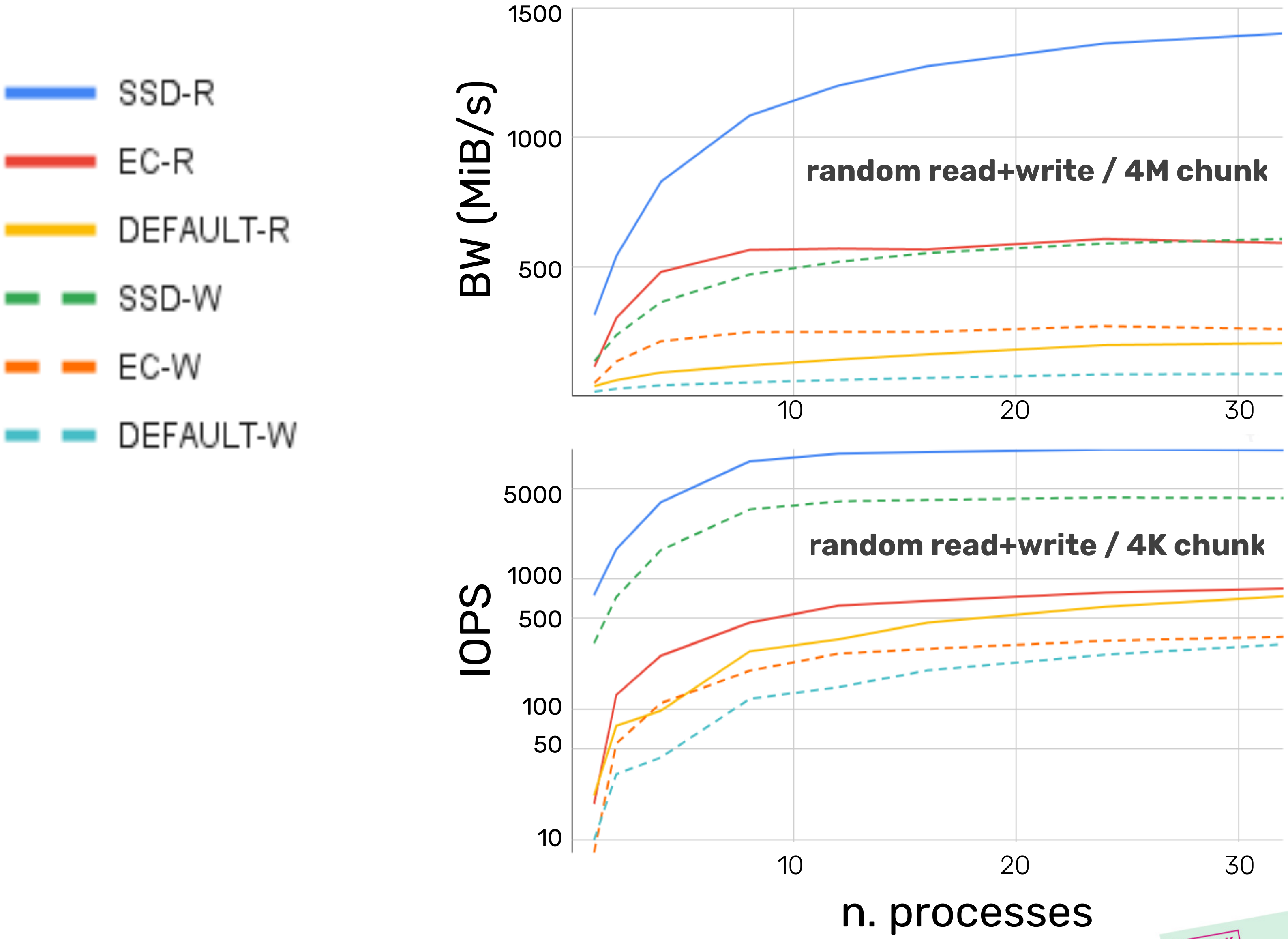


# From Ceph to OpenStack



| | Block store (cinder) | | | Object store | |
|---|---|---|---|---|---|
| volume ty | defau | fa st | capaci ty | swift / rclone / s3 / ... | |
| Interfac | rbd | rbd | rbd | rgw | rgw |
| Data Protecti | replica3 | replica3 | data: HDD EC6+4<br>metadata: SSD repl3 | EC6+4 | EC6+4 |
| Device Class | hdd | ssd | big | big | big |
| Devic | | SSD | SSD | | |

# Benchmarks

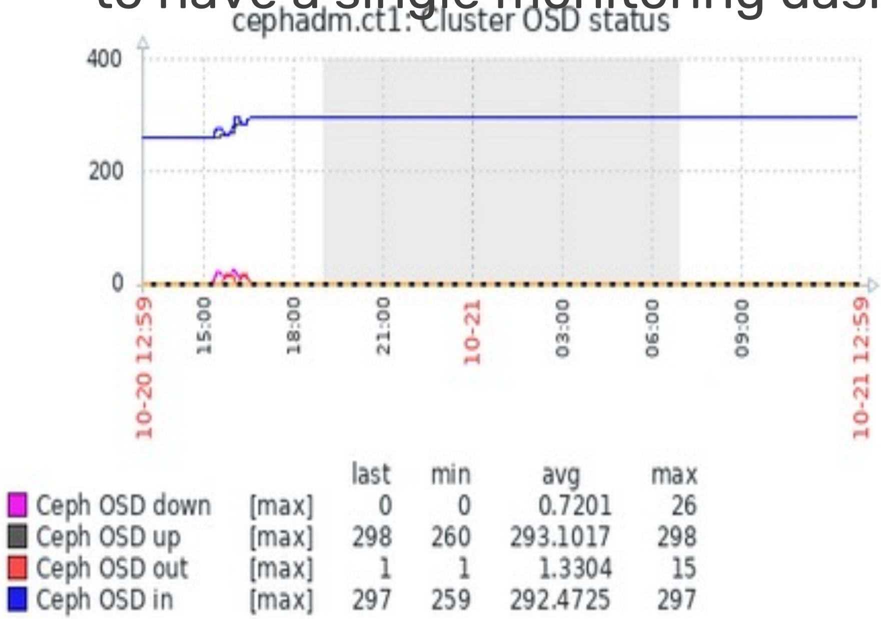

SSD-R
EC-R
DEFAULT-R
SSD-W
EC-W
DEFAULT-W
BW (MiB/s)
1500
1000
500
random read+write / 4M chunk
10
20
30
IOPS
5000
1000
500
100
50
10
random read+write / 4K chunk
10
20
30
n. processes

# Monitoring

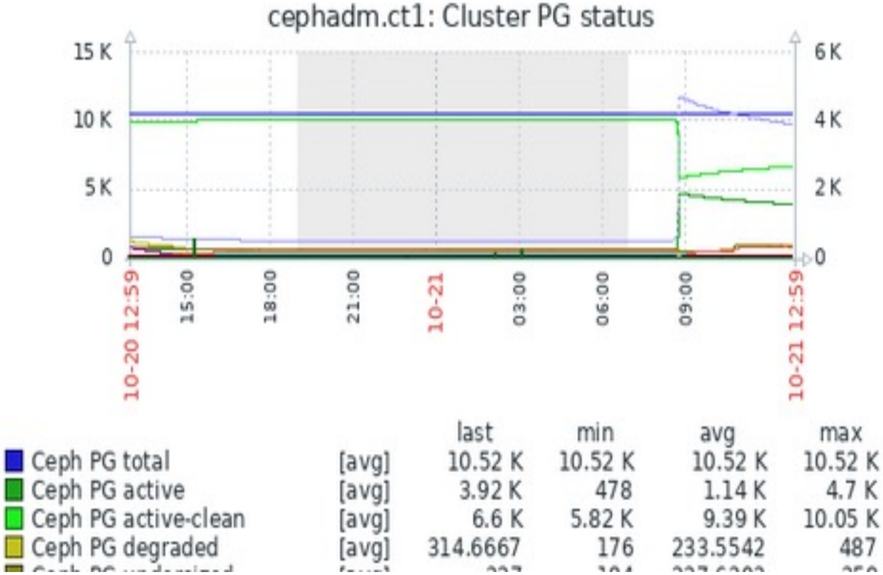
- Currently: Zabbix monitoring configured with custom Ansible scripts
- Ceph recently included native API to Zabbix and Grafana

  ➡️ We will integrate Ceph and Grafana,

  to have a single monitoring dashboard for all the cloud systems

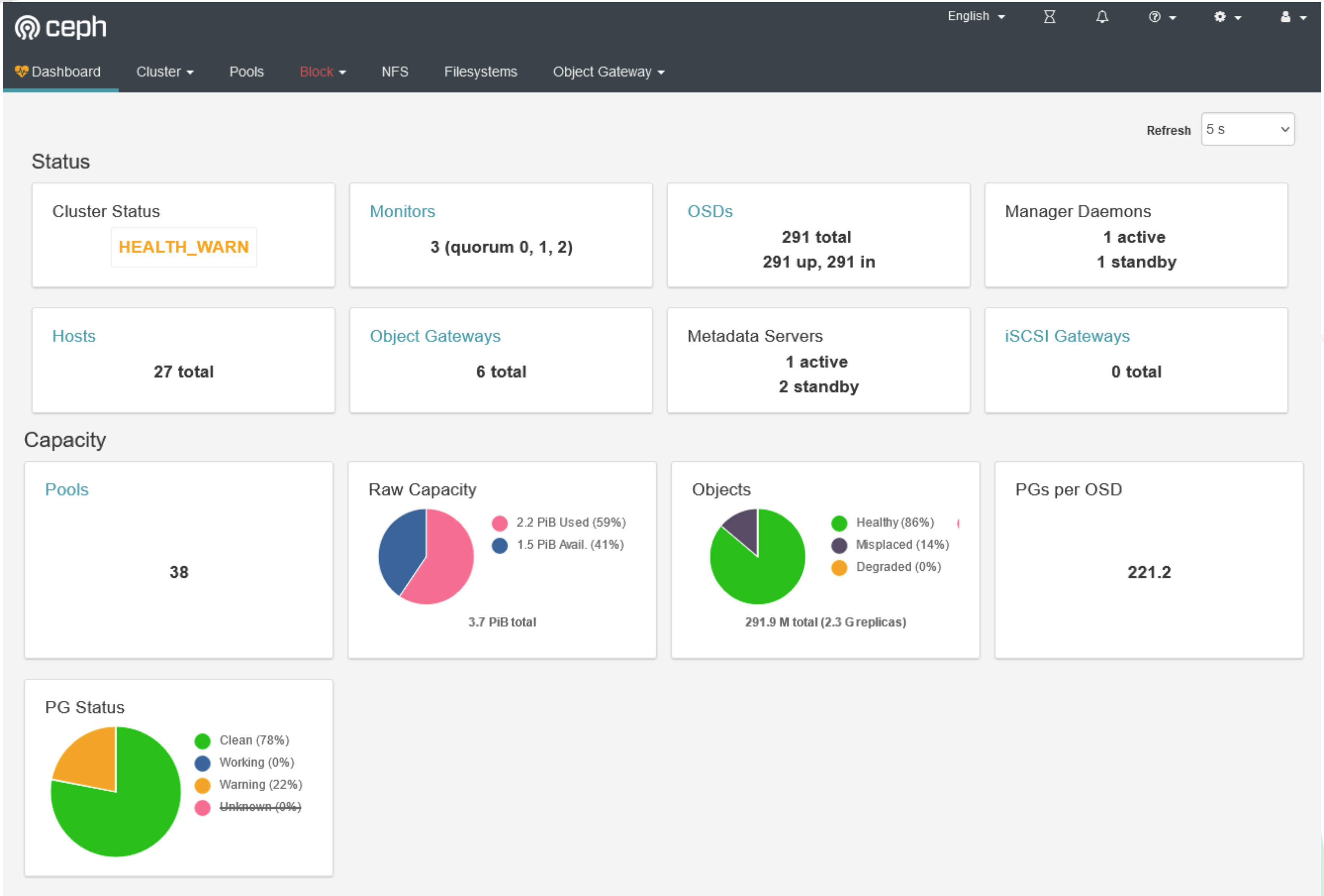


| | | last | min | avg | max |
|---|---|---|---|---|---|
| 🟪 Ceph OSD down | [max] | 0 | 0 | 0.7201 | 26 |
| ⬛ Ceph OSD up | [max] | 298 | 260 | 293.1017 | 298 |
| 🟥 Ceph OSD out | [max] | 1 | 1 | 1.3304 | 15 |
| 🟦 Ceph OSD in | [max] | 297 | 259 | 292.4725 | 297 |



| | | last | min | avg | max |
|---|---|---|---|---|---|
| 🟦 Ceph PG total | [avg] | 10.52 K | 10.52 K | 10.52 K | 10.52 K |
| 🟩 Ceph PG active | [avg] | 3.92 K | 478 | 1.14 K | 4.7 K |
| 🟩 Ceph PG active-clean | [avg] | 6.6 K | 5.82 K | 9.39 K | 10.05 K |
| 🟨 Ceph PG degraded | [avg] | 314.6667 | 176 | 233.5542 | 487 |

# Monitoring - Ceph dashboard



Refresh 5 s

Status

Cluster Status
HEALTH_WARN

Monitors
3 (quorum 0, 1, 2)

OSDs
291 total
291 up, 291 in

Manager Daemons
1 active
1 standby

Hosts
27 total

Object Gateways
6 total

Metadata Servers
1 active
2 standby

iSCSI Gateways
0 total

Capacity

Pools
38

Raw Capacity
● 2.2 PiB Used (59%)
● 1.5 PiB Avail. (41%)
3.7 PiB total

Objects
● Healthy (86%)
● Misplaced (14%)
● Degraded (0%)
291.9 M total (2.3 G replicas)

PGs per OSD
221.2

PG Status
● Clean (78%)
● Working (0%)
● Warning (22%)
● ~~Unknown (0%)~~

# Ceph management – a little cookbook

- Enable “balancer”: can also tune activity window and threshold

```
ceph balancer on ; ceph balancer mode upmap ; # default mode
```

- Occasionally, “balancer” fails and one OSD may “drift”
  - Adjust OSD weight, let it go for ~hour, set it back

```
ceph osd crush reweight osd.XXX 0.9
```

- Reduce impact of rebalancing/recovering operations:

```
ceph tell 'osd.*' injectargs --osd-recovery-sleep=0.1
      --osd-max-backfills=1 --osd-recovery-max-active=3
      --osd-recovery-max-single-start=1
```

  - https://docs.ceph.com/en/latest/rados/configuration/osd-config-ref/
  - ignore message stating restart is required

- Throttle data movement when adding servers

```
ceph balancer off
```

  - add servers, move OSD disks, do whatever…
    (see https://github.com/digitalocean/pgremapper or https://github.com/cernceph/ceph-scripts.git)

```
./tools/upmap/upmap-remapped.py | sh
```

  - re-run, to force all “remapped” PGs in their current position

```
ceph balancer on ; # let balancer gradually move PGs
```

  - at any time, can pause data movement by “upmap-remapped.py”

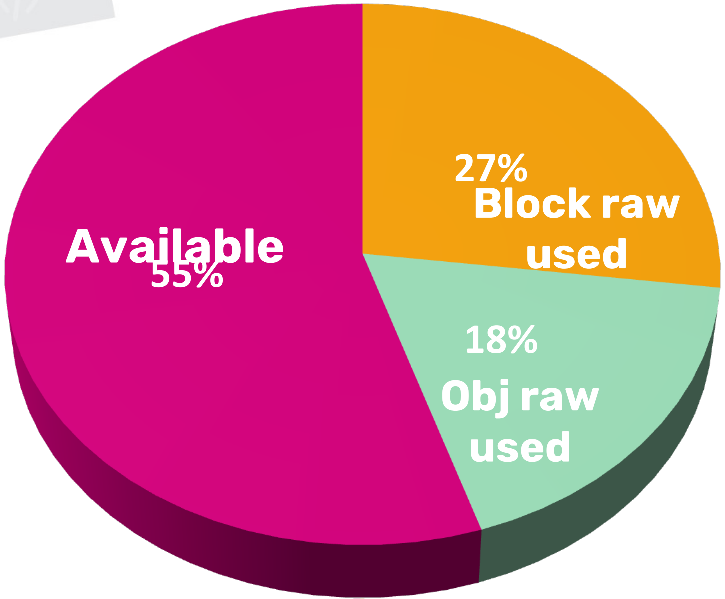
# Object store clients

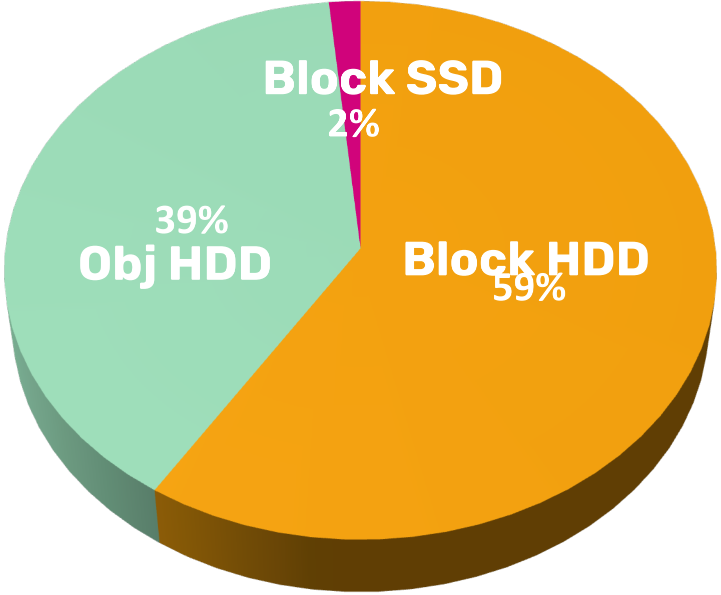
- rclone
  - swift endpoint
  - cli provides functions equivalent to rsync, cp, mv, ls, ncdu, tree, rm, etc.
  - **rclone mount**: mounts Object store as a disk on many systems
  - (optional) **server side encryption**
  - Works on Windows and Linux
  - (Experimental) GUI available

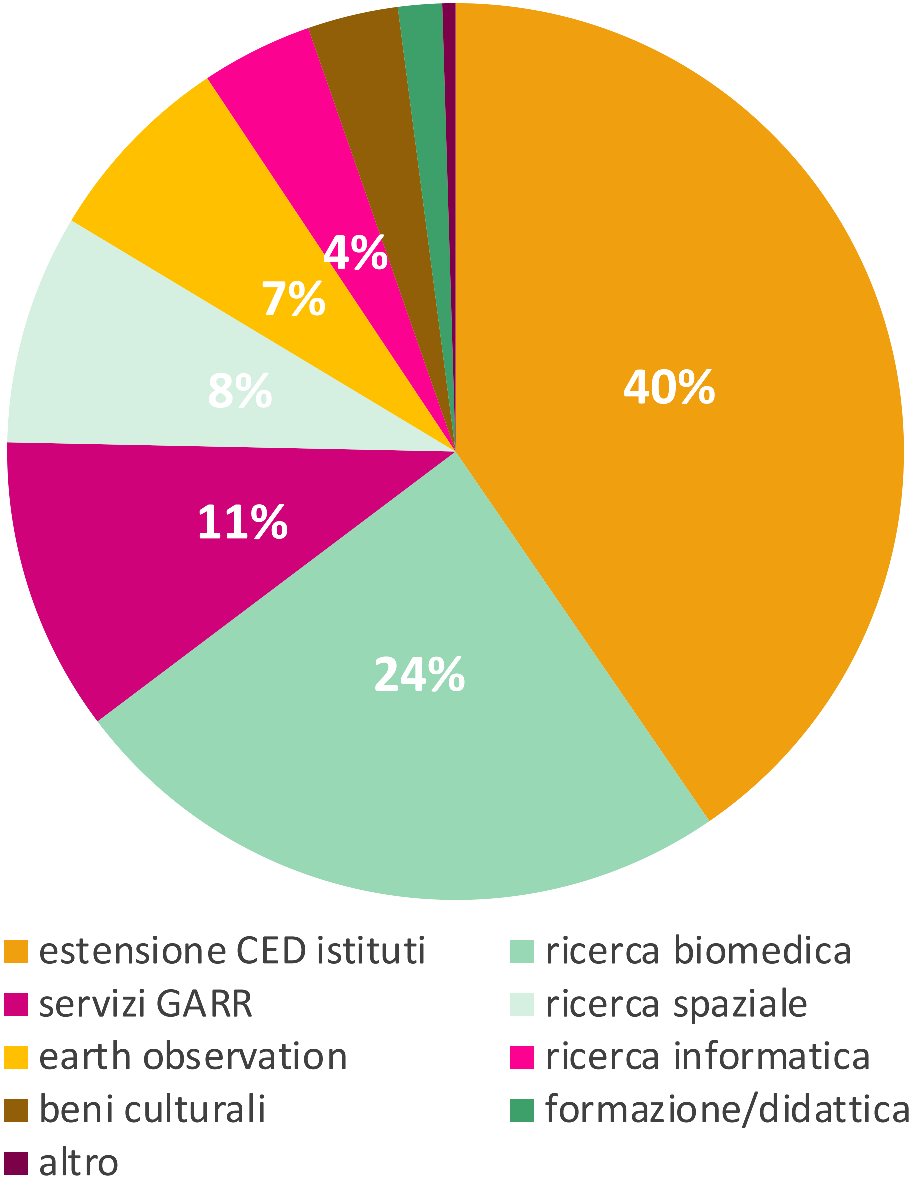
- s3cmd
  - s3 endpoint
  - client for uploading, retrieving and managing data in s3 bucket

- s3fs
  - s3 endpoint
  - large subset of POSIX including reading/writing files, directories, symlinks, mode, uid/gid, and extended attributes
  - optional server side encryption

# GARR Cloud storage utilization







# Conclusions

- GARR Cloud uses Ceph since 2015 with satisfaction
  - Great direct and indirect support from ceph-users@ceph.io
  - Upgraded seamlessly from Ceph v0.9 (Hammer) to v16 (Octopus)
  - ~10 PB raw storage managed
  - 1.5 PB net / 4.5 PB raw currently used / requested by 250 active users
  - Hot data migration to the new hardware infrastructure done
    - Not a single bit lost; performance degradation during migrations handled
  - Complete configuration of new hardware by end of November

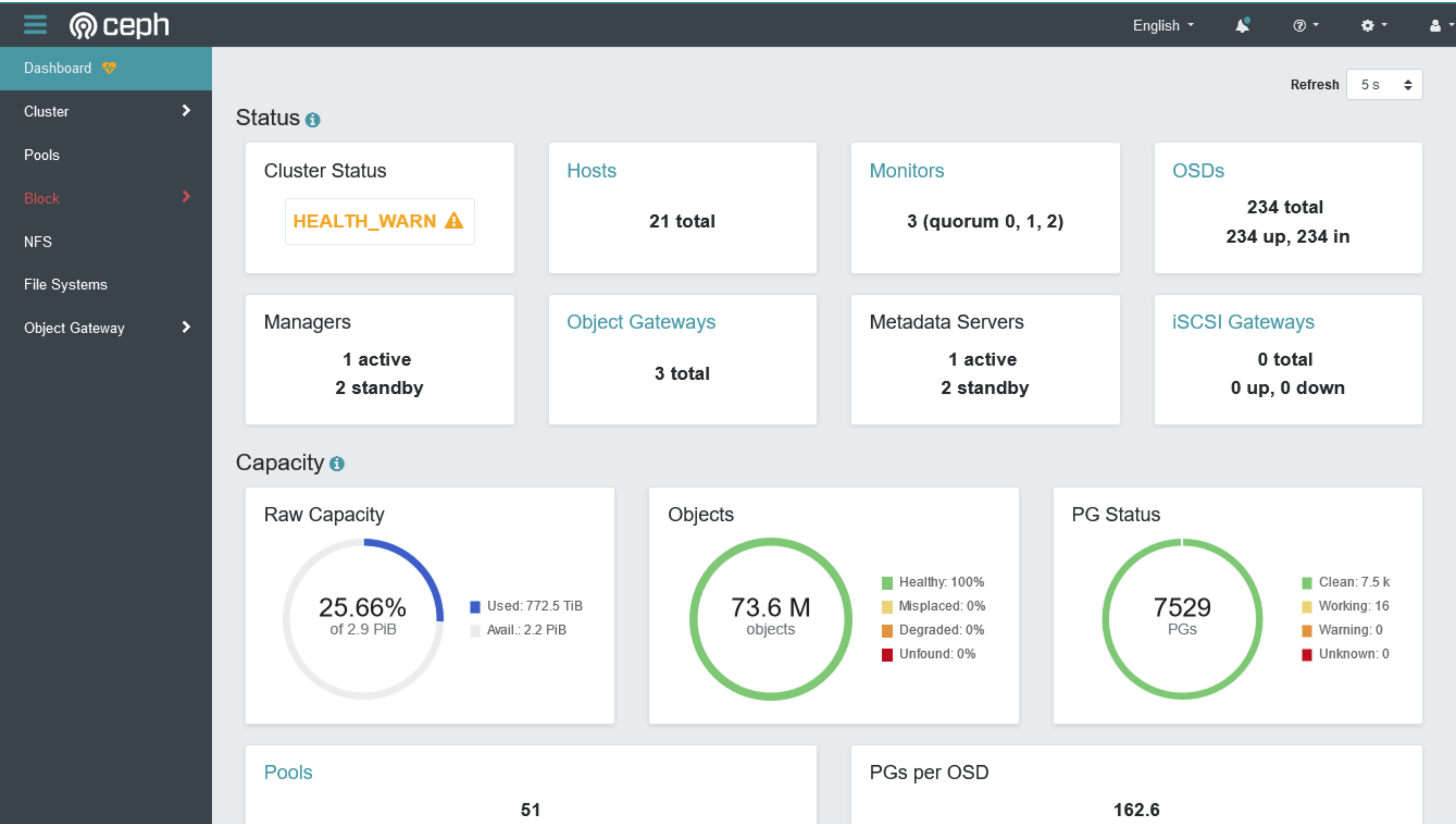
- Next steps
  - Complete CT⇔PA mirroring of Object Store and Glance (Cloud images) pools
  - Implement Ceph File System provisioning (OpenStack Manila)

Questions
?

photo © Gabriele Coll

# Backup

# Monitoring (Pacific: PA1)



Refresh 5 s

Status

Cluster Status
HEALTH_WARN ⚠

Hosts
21 total

Monitors
3 (quorum 0, 1, 2)

OSDs
234 total
234 up, 234 in

Managers
1 active
2 standby

Object Gateways
3 total

Metadata Servers
1 active
2 standby

iSCSI Gateways
0 total
0 up, 0 down

Capacity

Raw Capacity
25.66%
of 2.9 PiB
🟦 Used: 772.5 TiB
⬜ Avail.: 2.2 PiB

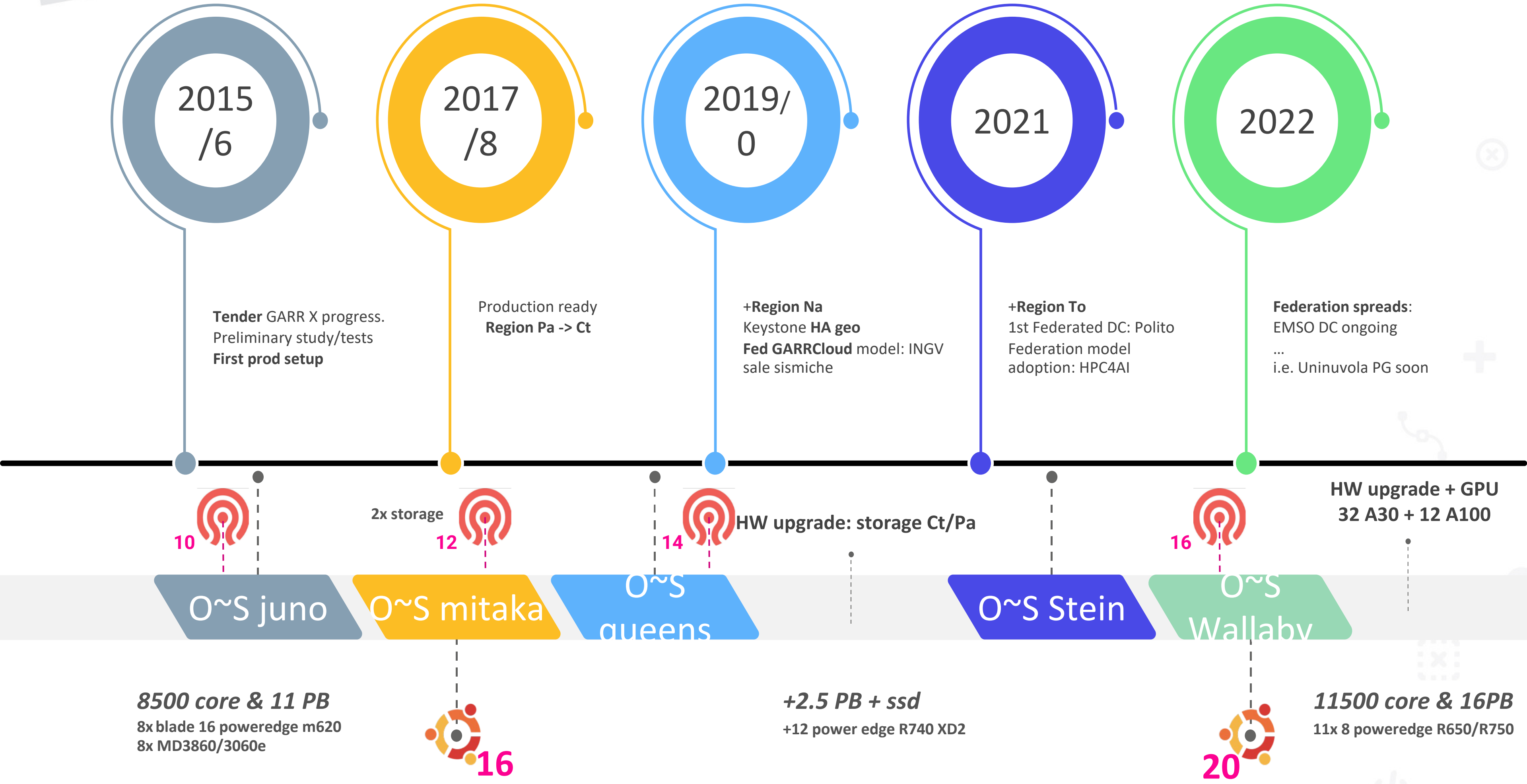
Objects
73.6 M
objects
🟩 Healthy: 100%
🟨 Misplaced: 0%
🟧 Degraded: 0%
🟥 Unfound: 0%

PG Status
7529
PGs
🟩 Clean: 7.5 k
🟨 Working: 16
🟧 Warning: 0
🟥 Unknown: 0

Pools
51

PGs per OSD
162.6

# cloud evolution timeline



2015/6

**Tender** GARR X progress.
Preliminary study/tests
**First prod setup**

2017/8

Production ready
**Region Pa -> Ct**

2019/0

**+Region Na**
Keystone **HA geo**
**Fed GARRCloud** model: INGV sale sismiche

2021

**+Region To**
1st Federated DC: Polito
Federation model adoption: HPC4AI

2022

**Federation spreads**:
EMSO DC ongoing
…
i.e. Uninuvola PG soon

**2x storage**

**HW upgrade: storage Ct/Pa**

**HW upgrade + GPU**
**32 A30 + 12 A100**

***8500 core & 11 PB***
**8x blade 16 poweredge m620**
**8x MD3860/3060e**

***+2.5 PB + ssd***
**+12 power edge R740 XD2**

***11500 core & 16PB***
**11x 8 poweredge R650/R750**