

Ceph

Infrastructure Storage at CERN

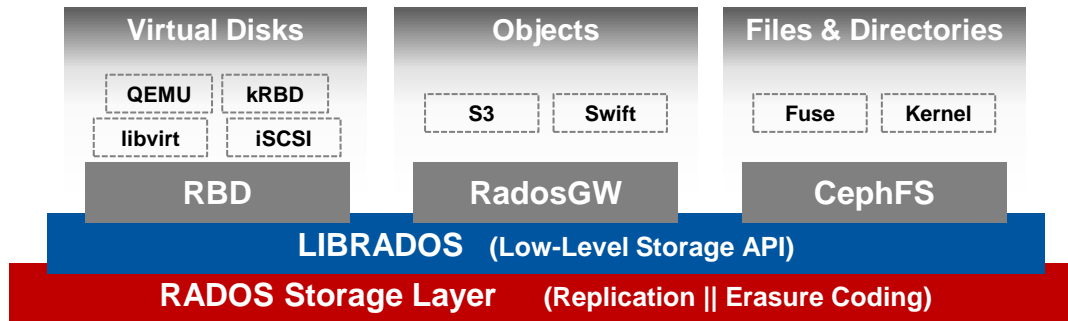
Enrico Bocchi
CERN IT, Storage

Pictet visit
27 September 2024



What is Ceph?

- Distributed Storage System, Open Source
- Reliable storage out of unreliable components:
 - Runs on commodity hardware (IP networks, HDDs/SSDs/NVMes)
 - Favors data consistency and correctness over performance and availability
- Elastic and self-healing:
 - Scale up or out online and under load (or similarly shrink)
 - Self-recovery from HW failures, res-establishing desired redundancy



What does Ceph do for us?

- Storage backbone underpinning CERN's IT Cloud and Services
 - Code repositories, Container Registries, GitOps, Agile Infrastructure
 - Document / Web Hosting
 - Monitoring: OpenSearch, Kafka, Grafana, InfluxDB, Kibana
 - Analytics: HTCondor, Slurm, Jupyter Notebooks, Spark
 - Virtualization of other Storage: NFS, AFS, CVMFS, ...

Application		Size (raw)	Clusters
Blocks	<i>HDD, 3x replica</i>	25.1 PB	5
	OS Cinder/Glance <i>Flash, EC 4+2</i>	976 TB	2
File System	<i>HDD, 3x replica</i>	13.4 PB	5
	OS Manila, K8s/OKD, HPC <i>Flash, 3x replica</i>	1.7 PB	4
Objects	<i>HDD, EC 4+2</i>	28.2 PB	2
	S3, Swift, Backups <i>Multi-site, EC 4+2</i>	3.6 PB	1

What does Ceph do for us?

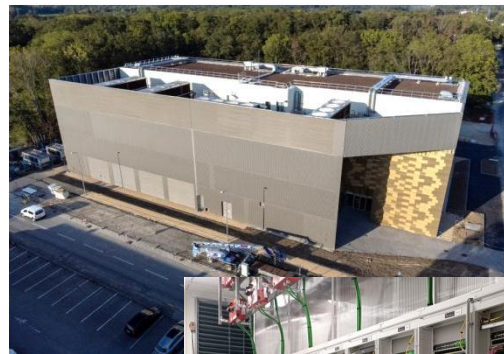


Service History

- 2013: 300TB proof of concept (replica 4!) → 1 cluster, 3 PB in production for RBD
- 2016: 3PB to 6PB expansion with no downtime
- 2018: S3 + CephFS in production
- 2020: S3 Backup cluster in 2nd location
- 2021: RBD Storage Availability Zones
- 2022: CephFS cluster physical move with no downtime
- 2023: KernelRBD in production
- 2024: New Datacenter!

- ✓ 19 production clusters
 “don’t put all your eggs in the same basket”
- ✓ 5 additional clusters in new datacenter
- ✓ Exotic cluster configurations
 - Cross-DC *stretch* clusters
 - S3 multi-site objects replication

Under
Evaluation



Integration with OpenStack

- OpenStack is the entry point for compute and storage resources:

- Ceph Blocks → Cinder volumes + Glance images for VMs
- S3 Objects → Keystone as vault for authentication keypairs
- CephFS → Manila FileShares

- IaaS components are self-service to end-users

- Example of Block storage provisioning
- Quota is subject to our (Cloud+Ceph) approval, which is also an opportunity to guide users

Volume Type	QoS	Pool Type	AZs
standard	80MB/s, 100 IOps	3x Replicas	3 Zones
io1	120MB/s, 500 IOps		
io2	300MB/s, 1000 IOps	EC 4+2	
io3	300MB/s, 5 IO per GB (min 500, max 2000)	Full-Flash	-

The screenshot shows the OpenStack dashboard interface. On the left, a navigation menu includes 'Details', 'Compute', 'Volumes' (highlighted), 'File Shares', 'Object Storage', 'Network', and 'Load Balancer'. The main content area is divided into two panels:

- QoS** (Quality of Service): A table listing volume types and their QoS parameters.

Volume Type	QoS
cp1	
cpio1	
cpio2	
cpio3	
io1	
io2	
io3	
standard	
- Quota**: A table showing resource limits for different volume types.

Number of Volumes	Space in Gigabytes
12	5000
12	5000
1	10
1	10
10	10000
0	0
5	5000
50	10000

Below these panels is a **Backups** section with a form for specifying backup resources:

If backups are required, please provide the desired amount of resources below:

Number of Backups:

Space in Gigabytes:

Backup Quota

A few words on Hardware and Network

- 2 main hardware types for any of blocks, file system**, and objects:

- HDD server:

- Frontend with a handful of SSD devices (OS, Ceph journals), 25Gbps NIC, SAS controller
- 2x JBOD with 24 enterprise CMR HDDs

- Full flash server:

- 2U node with 10x NVMe (was SATA SSDs)

- Cores and memory depend on number of drives



- Network:

- Ceph supports cluster VS public (i.e., anything else) networks, IPv4 or IPv6 + TCP
- It may be network hungry when doing major rebalancing or recovery operations

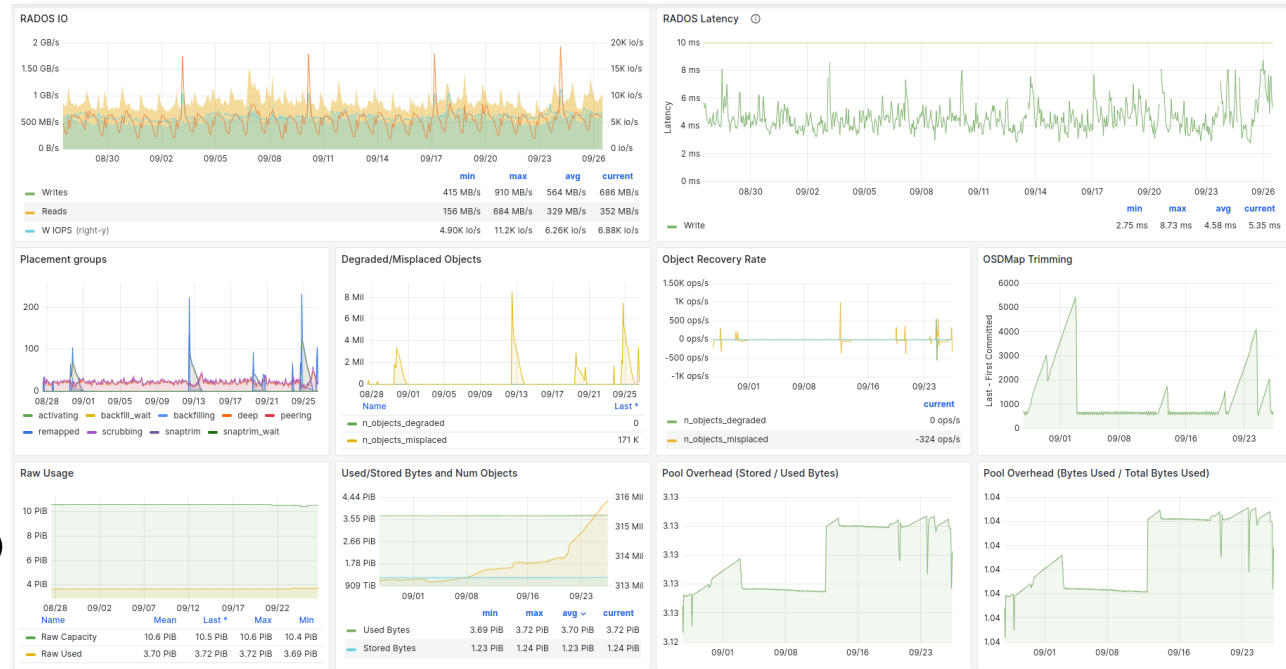
** CephFS at scale needs extra care:

- Metadata is stored on a dedicated pool, which loves to be on flash drives
- Metadata server (MDS) requires memory: Consider 64+ GB per MDS, can scale out horizontally

A few words on Monitoring

Metrics + Logs

- Prometheus Node exporter
- Ceph Prometheus module
- OpenStack exporter for metrics integration
- Prometheus local – last 48h
- Thanos store (on S3) for long-term archival
- Grafana for visualization
- Several homegrown scripts remain for custom metrics (latency, PSI, S3 checks, ...)
- Logs?
Fluentbit, Kafka, Logstash, OpenSearch



30 days on our main Block Storage cluster

Learn more about Ceph

- Website: ceph.io
- Mailing List: ceph-users@ceph.io
- Community [Google Calendar](#) – Monthly User+Dev Meetup + Tech Talks
- Cephalocon – Flagship yearly conference at CERN in December!

Registrations
Open!

cephalocon
ceph

Register Attend Sponsor Program Contact Us View

4-5 DECEMBER 2024
CERN
GENEVA, SWITZERLAND
#Cephalocon

SUBMIT A PROPOSAL SPONSOR
REGISTER NOW

VENUE INFORMATION

CERN SCIENCE GATEWAY
Espl. des Particules 1
1217 Meyrin, Switzerland

Running 4-5 December, Cephalocon brings together a global community of operators, developers, and researchers to celebrate Ceph, the open source distributed storage system designed to deliver excellent performance, reliability, and scalability.

Discussion



Enrico Bocchi
enrico.bocchi@cern.ch