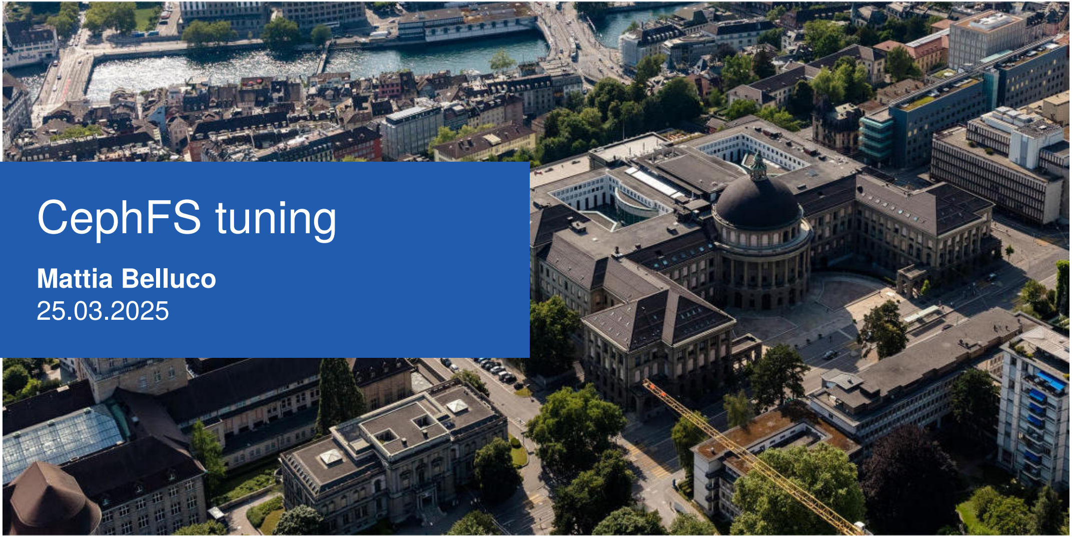
ETH zürich

Scientific IT Services



# CephFS tuning

**Mattia Belluco**

25.03.2025

# Outline

1. CephFS @ SIS
2. CephFS best practices
3. CephFS in production
4. What we would do differently

# Outline

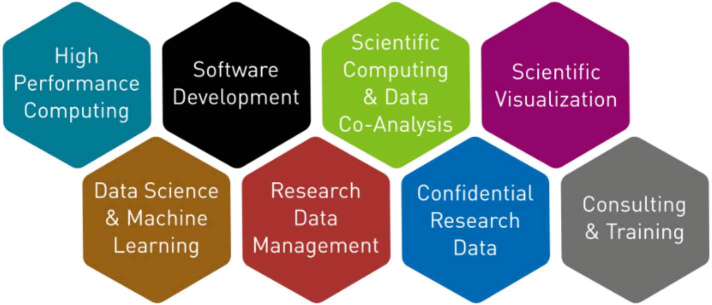
1. CephFS @ SIS

2. CephFS best practices

3. CephFS in production

4. What we would do differently

# SIS: Scientific IT Services


High Performance Computing
Software Development
Scientific Computing & Data Co-Analysis
Scientific Visualization
Data Science & Machine Learning
Research Data Management
Confidential Research Data
Consulting & Training

- A section of ETH Zürich IT Services
- Composed of ~40 experts in various areas of scientific computing
- Background in different areas of science

# CephFS context

## LeonhardMed Trusted Research Environment

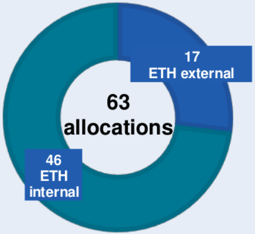
**> 507** researchers

**> 3.2 PB** secure data storage

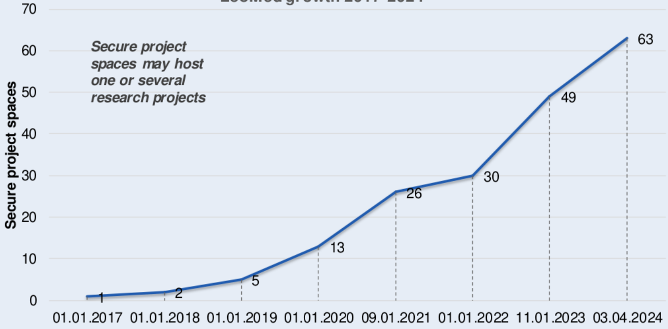
**> 10** Data Providers (DPs)
Swiss (e.g. USZ, EOC) and international (UCSF)

**> 1500** Data Transfer Requests from DPs in active reseach projects

**Leomed Customers Affiliation (Eth-internal Or External)**



**LeoMed growth 2017-2024**



*Secure project spaces may host one or several research projects*

# Outline

1. CephFS @ SIS

2. CephFS best practices

3. CephFS in production

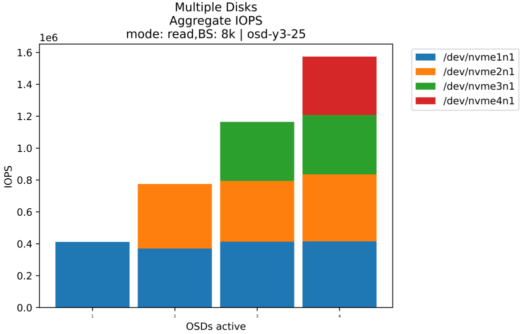
4. What we would do differently

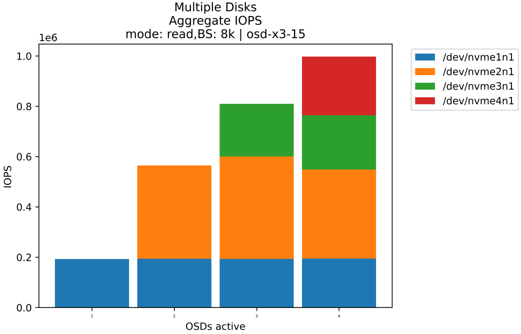
# Ceph(FS) guidelines we followed

- If using spindles, separated WAL+DB on fast devices
- Metadata pool on fast devices
- Filesystem root (default data pool) on a replicated (fast?!) pool
- Evaluate the selection of the EC algorithm (nice presentations at Cephalocon 2024)
- Validate hardware components (slow disks, faulty NICs w/ high error rates)

# Always validate your hardware performance

On one node a NVME device seems to be outperforming the other:





Looking closer at the Y axis it turns out **the exact opposite is true!**

# Outline

1. CephFS @ SIS

2. CephFS best practices

3. CephFS in production

4. What we would do differently

# CephFS in production

## Where synthetic benchmarks go to die

- System benchmarked with the default Pacific releases settings and 1 active MDS
- More than 1 PB of data migrated from the previous storage system
- Gradual ramp up by migrating individual projects to the new system
- It became soon clear that a single MDS daemon could not cope with the amount of requests originating from the clients.

# Some tuning needed

- Incrementally increase `max_mds` to help the metadata service to cope with the requests, up to the current value of `7`
- After the first increase to 3 we noticed a sudden increase of the `Req/s` counters that we later realized was due to the mds balancer exporting subtrees.
- We pinned the most active project directories to dedicated MDS demons (later we pinned them all).

# Cache configuration

Doc page: https://docs.ceph.com/en/pacific/cephfs/cache-configuration/

- Lowered `mds_max_caps_per_client` to 524288 from 1000000
- Incrementally increase `mds_cache_memory_limit` up to the current value of `110 GB`
- Increase the mds `mds_cache_reservation` from 5% to 10% (i.e. `0.10`)
- Lowered the `mds_cache_threshold` value from `1.5` to `1.1`
- Lowered `mds_cache_trim_decay_rate` to 0.8 from 1.0
- Increased `mds_cache_trim_threshold` to 393216 from 256K for Quincy and 64K for Pacific

# Throttle deletion operation

Issue: deletion of millions of files caused the MDS to report being behind on trimming, eventually causing slow metadata ops.

Fix:

- Decrease `mds_max_purge_files` to `64`
- Decrease `mds_max_purge_ops_per_pg` to `0.500000`

Deletes would take longer but without ill effects on the MDS demon.

# Adjust MDS Recall

- `mds_recall_max_caps` to 20000 (from 5k in Pacific and 30k in Quincy)
- `mds_recall_max_decay_rate` to 2.000000 (from 2.5 Pacific and 1.5 in Quincy)
- `mds_recall_max_decay_threshold` to 65536 (from 16K in Pacific and 128K in Quincy)
- `mds_recall_global_max_decay_threshold` to 262144 (from 64K in Pacific and 128K in Quincy)
- `mds_recall_warning_threshold` to 131072 (from 32K in Pacific and 256K in Quincy)

# MDS failover

When an MDS demon fails in a busy system like ours, the newly MDS assigned to the failed rank needs more time then the default value of 60 seconds to replay the journal and join the cluster. After several attempts `mds_beacon_grace` is set to `300` seconds

# CephFS production

## Configuration

| | | |
|---|---|---|
| advanced | mds bal interval | 0 |
| advanced | mds beacon grace | 300.000000 |
| basic | mds cache memory limit | 110000000000 |
| advanced | mds cache reservation | 0.100000 |
| advanced | mds cache trim decay rate | 0.800000 |
| advanced | mds cache trim threshold | 393216 |
| advanced | mds health cache threshold | 1.100000 |
| advanced | mds log max segments | 128 |
| advanced | mds max caps per client | 524288 |
| advanced | mds max purge files | 64 |
| advanced | mds max purge ops per pg | 0.500000 |
| advanced | mds recall global max decay threshold | 262144 |
| advanced | mds recall max caps | 20000 |
| advanced | mds recall max decay rate | 2.000000 |
| advanced | mds recall max decay threshold | 65536 |
| advanced | mds recall warning threshold | 131072 |

# Outline

1. CephFS @ SIS
2. CephFS best practices
3. CephFS in production
4. What we would do differently

# CephFS data pool on a replicated flash pool

Replication required since Nautilus because of the tiny objects that are created on the root of the filesystem with backtrace information (used in Disaster Recovery) and hardlinks references.

Problem:

- Currently around 560 M inodes in our 480 OSDs filesystem
- 3.5M tiny objects per disk that needs to be rebalanced if a drive breaks.

In our system it takes **as much time** to rebalance the default "empty" cephfs datapool as it tkes to rebalance the actual data.

# Evaluate the selection of the EC algorithm with more care

Presentations at Cephalocon 2024 by Jamie Pryde from IBM shows how performance is highly dependent on the chosen Erasure Coding algorithm and debunk the belief that ISA-L only works only on Intel CPUs.

Reference:
Erasure Coding: 5 Ways to Split a Squid
https://www.youtube.com/watch?v=aM8sJgDD-x4

# Thanks for your attention

Questions?

ETH zürich

Mattia Belluco
Cloud Architect
mattia.belluco@id.ethz.ch

ETH Zurich
IT Services
OCT G 35
Binzmühlestrasse 130
8092 Zürich, Switzerland
https://sis.id.ethz.ch