Ceph in 2023 & Beyond

HEPiX Autumn 2023 Workshop
October 18, 2023

Dan van der Ster
Ceph Executive Council / CTO Clyso GmbH

About Me

- University of Victoria - 1998:
    - B.Eng in Computer Engineering @ UVic
    - PhD in Grid Computing @ UVic – ***Supervisor Dr. Randall J. Sobie***
- CERN - 2008
    - Grid Group: ATLAS Distributed Analysis Dev and Coordinator 2008-2012
    - Storage Group: AFS, CVMFS, Ceph Service Manager 2013-2022
    - Governance Group: Chief IT Architect 2022-2023
    - Sabbatical Leave 2023-present
- Ceph Open Source Project - 2013:
    - Ceph Foundation Board Member 2015-present
    - Ceph Executive Council 2021-present
- Clyso GmbH - 2023
    - CTO – leading North American expansion

Outline

- Brief Introduction to Ceph
- Recent Developments
- Ceph Community News
- What I’m working on

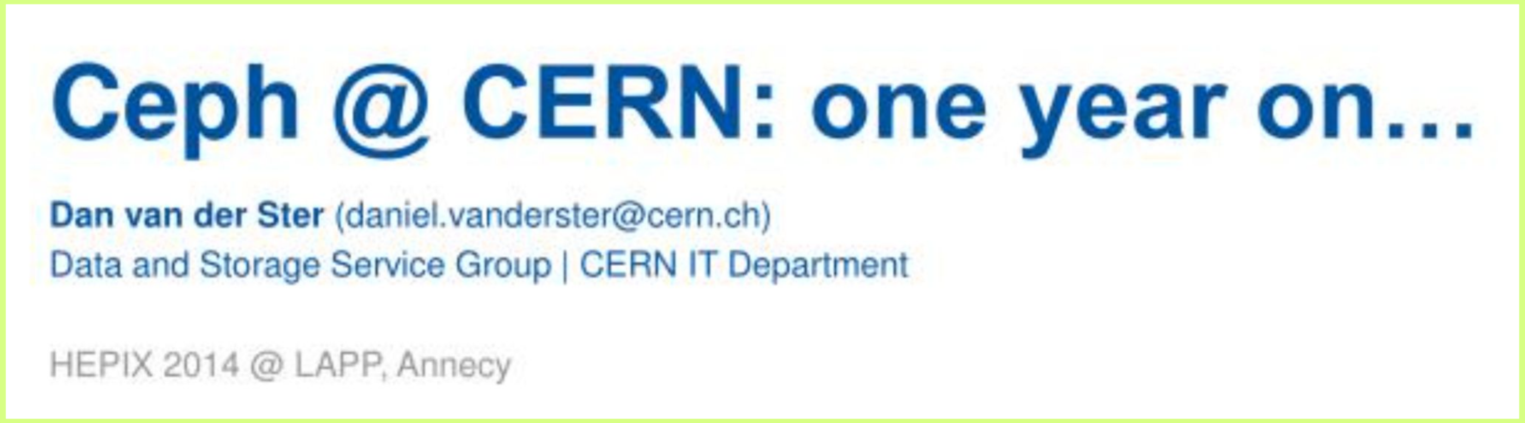
Introduction to Ceph

- How many of you know Ceph? *operate* Ceph? *like/dislike* Ceph?
- Built upon a Reliable Autonomic Distributed Object Store: **RADOS**
- Objects are distributed pseudorandomly using **CRUSH**
- End result:
    - Enterprise-quality Block, File, and Object storage using commodity hardware
    - Scalable, reliable, organic technology backing much of the world's cloud infrastructures
    - Open Source Software – the **Linux of Storage**

History of Ceph

- 2007 - Sage Weil’s PhD on CRUSH and CephFS
- 2011 - Inktank startup founded to commercialize Ceph
- *2013 - CERN started using Ceph*
- 2014 - Inktank acquired by Red Hat
- *2014 - Dan presented Ceph@CERN: One year on.. At HEPIX LAPP*
- 2018 - Creation of the Ceph Foundation
- 2019 - Red Hat acquired by IBM
- 2023 - Ceph team reassigned from RH to IBM

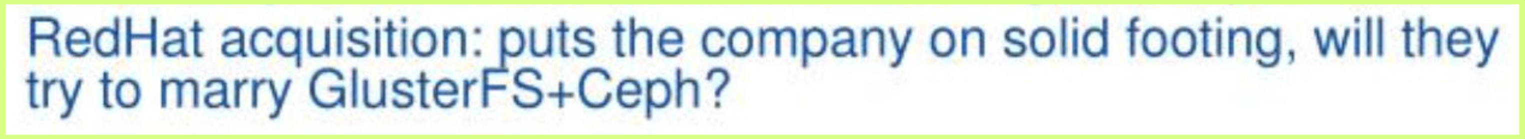
History of Ceph



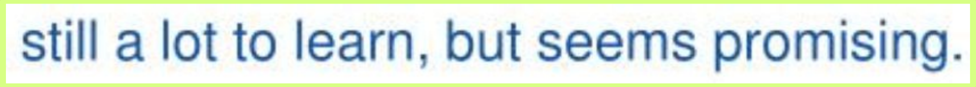
Ceph @ CERN: one year on…

Dan van der Ster (daniel.vandersterr@cern.ch)
Data and Storage Service Group | CERN IT Department

HEPIX 2014 @ LAPP, Annecy



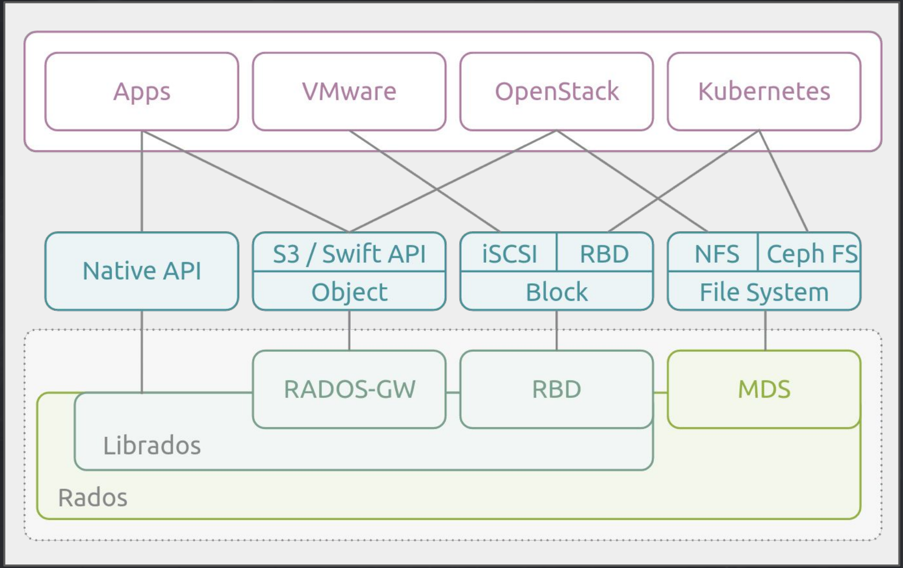
RedHat acquisition: puts the company on solid footing, will they try to marry GlusterFS+Ceph?



still a lot to learn, but seems promising.

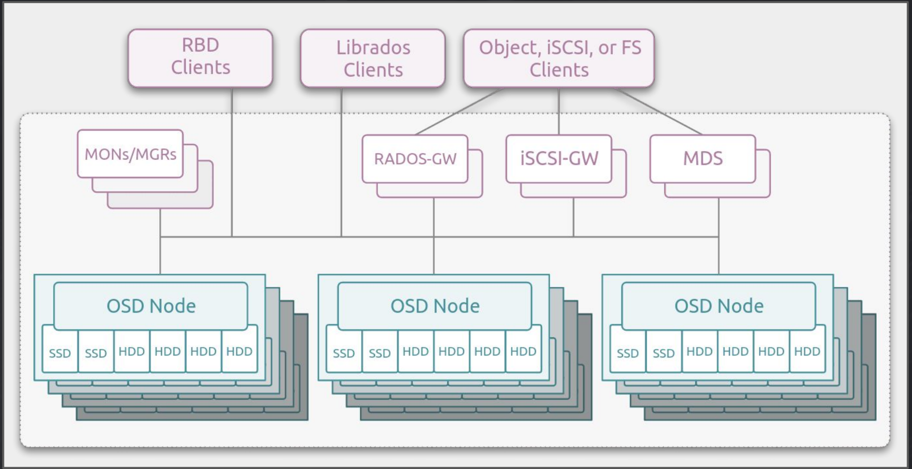
Ceph Architecture

- **RADOS:** low-level object store
- **RBD**: virtual block devices e.g. /dev/vdb attached to your VM
- **CephFS**: a shared network file system, mounted like NFS/AFS/...
- **S3**: HTTP-like object store, GET/PUT, AWS compatible.
- **Integrations:** OpenStack (Volumes, Shares, Object), Kubernetes (PVCs, Rook), ...



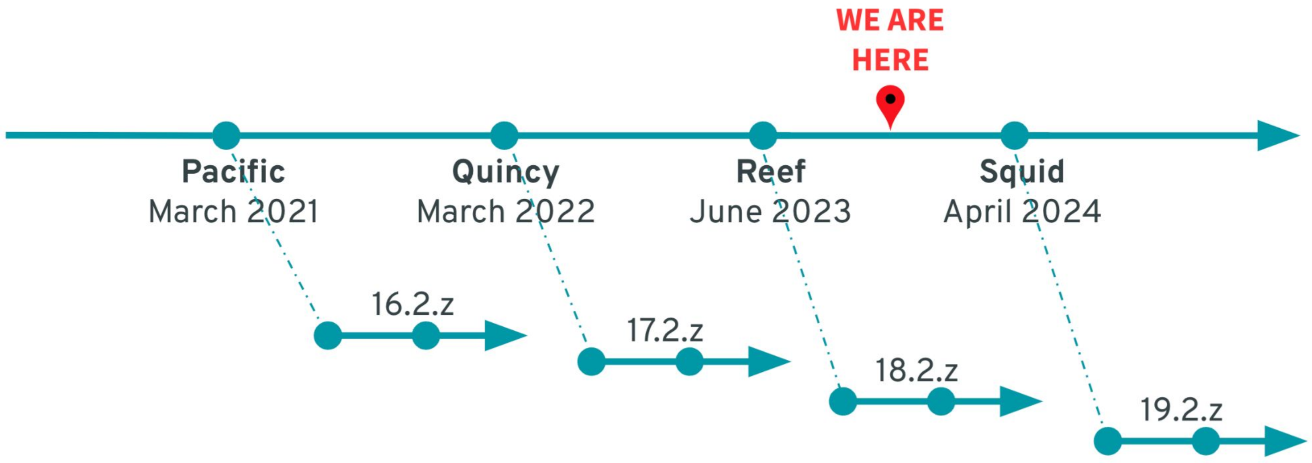
Ceph Components

- **OSDs (disks/NVMes)**
  - 4-8GB RAM per device
  - BlueStore+RocksDB on-disk format
- **MON/MGR**
  - Central cluster maps, not in IO path
  - Smallish servers, Reliable via PAXOS
- **MDS (CephFS)**
  - Scale-out metadata, hot/cold standbys
  - O(100GB) RAM each, single threaded
- **RGW (S3)**
  - Scale-out S3-compatible gateways
  - Multi-region support



All built on commodity hardware

Ceph Software Releases



- Stable, named release every 12 months
- Backports for 2 releases
  - Pacific reaches EOL shortly after Reef is released
- Upgrade up to 2 releases at a time
  - Nautilus → Pacific, Pacific → Reef, Quincy → S

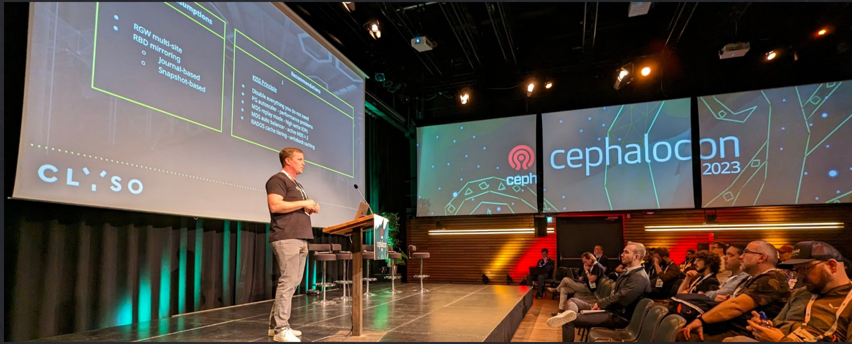
https://ceph.io

Reef v18 Highlights

- (Please don’t be underwhelmed – Ceph is stable software)
- **RADOS**: mem usage fixes, dist QoS with mclock, custom WAL, 4kB alloc units for BlueFS, read IO balancer
- **RBD**: NVMeoF target gateway, persistent wb cache, rbd-mirror ++
- **CephFS**: cephfs-top, fscrypt, stability ++
- **RGW**: rate limiting, SSE-S3, s3select, multisite replication ++
- **Dashboard**: 1-click OSD create, capacity planning, upgrades, S3 multisite, S3 policy admin

Ceph Community

- Ceph Foundation
    - 40 corporate + associate members
    - Supports neutral upstream development, testing, documentation, events, marketing
- Events:
    - Ceph Days 2023 - NYC, SoCal, India, Seoul, Vancouver
    - Cephalocon 2023 - Amsterdam
    - All talks recorded and shared on Youtube
- Securing the Foundation:
    - New tiers to secure the project’s future
    - Plans to invest in more infra, bigger events
- Technical Meetups:
    - Ceph Leadership Team + Component Weekly
    - Ceph Developer Monthly



What I’m working on

My Favourite Bugs

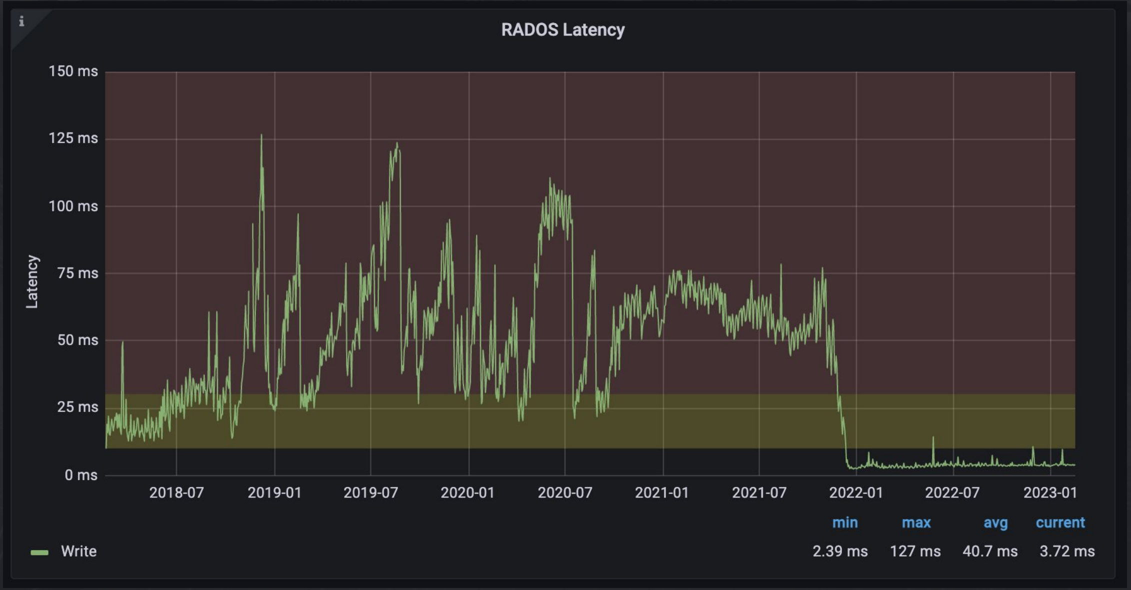
- Bug of the Year 2020: OSDMap LZ4 Corruptions
    - Symptom: Cluster-wide of OSD aborts with osdmap crc errors
    - Recovered the cluster by injecting an older valid osdmap
    - RCA: osdmaps had 4 flipped bits, caused by LZ4 which corrupted non-contiguous inputs in rare cases.
    - Solution: defrag ceph_buffers before compressing, and the OS upgraded its LZ4 library.
- Bug of the Year 2022: OSD PG Log "Dup" Bug
    - Symptom: For several months users reported OSDs consuming 100’s of GBs of RAM, even after restart. Mempool dumps showed huge allocations in the pg_log buffers.
    - RCA: pg splitting and merging violated the ordering of the duplicate op log, preventing trimming.
    - Solution: offline trim command for the OSD, and better online pg log management.

My Favourite Bugs

FIXED

- Bug of the Year 2020: OSDMap LZ4 Corruptions
    - Symptom: Cluster-wide of OSD aborts with osdmap crc errors
    - Recovered the cluster by injecting an older valid osdmap
    - RCA: osdmaps had 4 flipped bits, caused by LZ4 which corrupted non-contiguous inputs in rare cases.
    - Solution: defrag ceph_buffers before compressing, and the OS upgraded its LZ4 library.
- Bug of the Year 2022: OSD PG Log "Dup" Bug
    - Symptom: For several months users reported OSDs consuming 100’s of GBs of RAM, even after restart. Mempool dumps showed huge allocations in the pg_log buffers.
    - RCA: pg splitting and merging violated the ordering of the duplicate op log, preventing trimming.
    - Solution: offline trim command for the OSD, and better online pg log management.

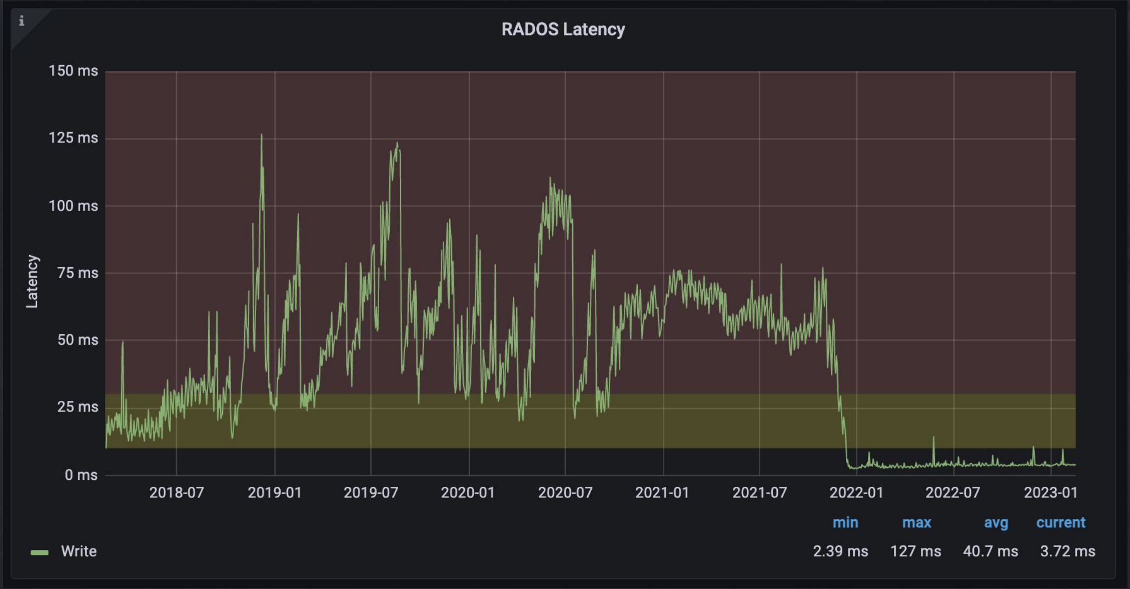
My Favourite Plot



	min	max	avg	current
Write	2.39 ms	127 ms	40.7 ms	3.72 ms

My Favourite Plot

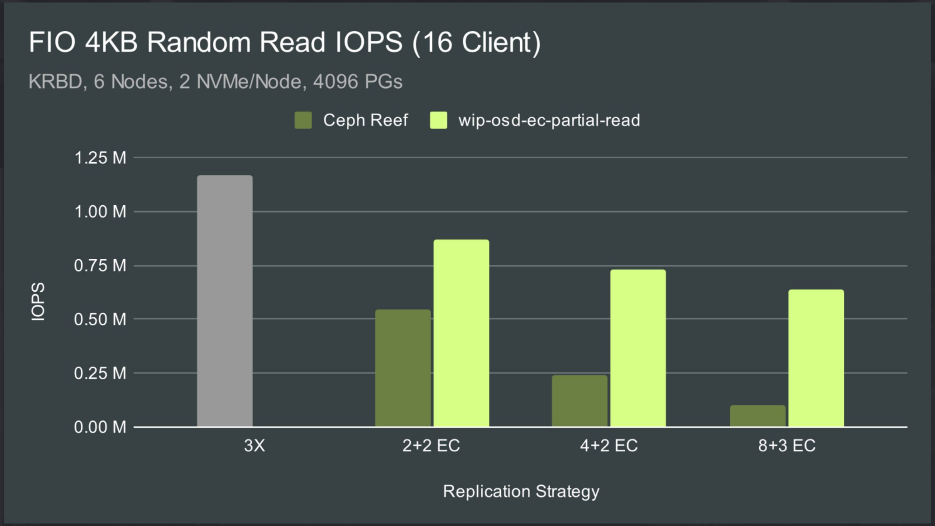
FIXED



Modern devices have a “media cache” which has a huge impact on BlueStore performance
Read ceph.com Hardware Recommendations re: disabling device writeback caches

My 2nd Favourite Plot

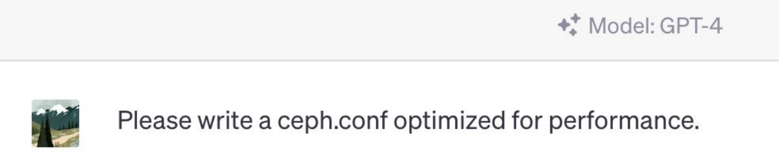
WIP



Potential 4x sped up IO path after workload analysis here at UVic!

Comparing Use-Cases

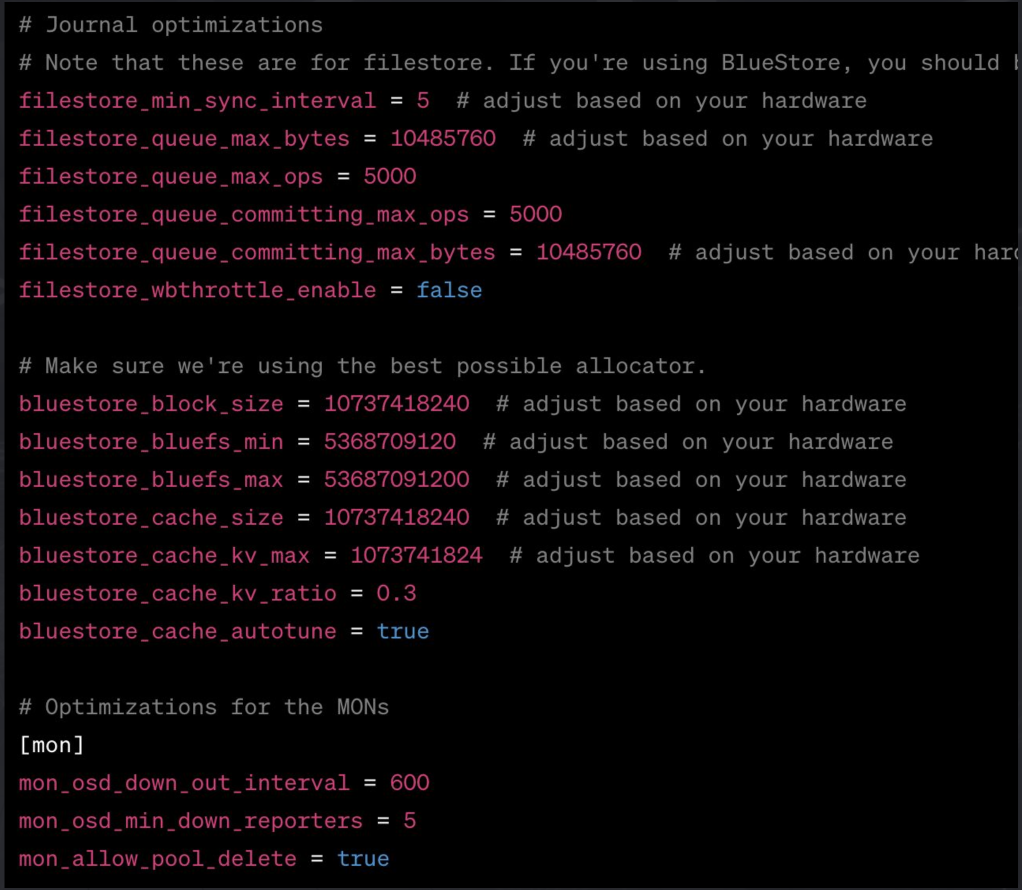
- CERN uses Ceph to back its cloud infrastructure: 100PB of block, S3, FS.
- In my new role I'm exposed to much more Ceph in very different envs:
  - Ranging from 10's of TB to multiples exabytes. Cluster in a closet to 100s of clusters globally.
  - "Microsoft/VMWare is too expensive". Moving to Proxmox+Ceph.
  - "Data is our product – We need full ownership of the platform."
  - "Ceph backs the things that make us money – if it's down we'll lose $$$ per minute"
  - "Xyz is too expensive, we're locked in ➜ FOSS Ceph is the best alternative we found"
- Lots and lots of successful uses out there – around 5 exabytes across thousands of clusters.
- But common themes – pain points – are emerging:
  - Ceph performance is not obvious – selecting hardware, NVMe, Crimson, multi-MDS, …
  - Ceph is *still* too difficult to understand and operate. ***#AI-OPS to the rescue?***



Please write a ceph.conf optimized for performance.

#AI-Ops ??

ChatGPT's recommendations vary between useless and *very dangerous*.

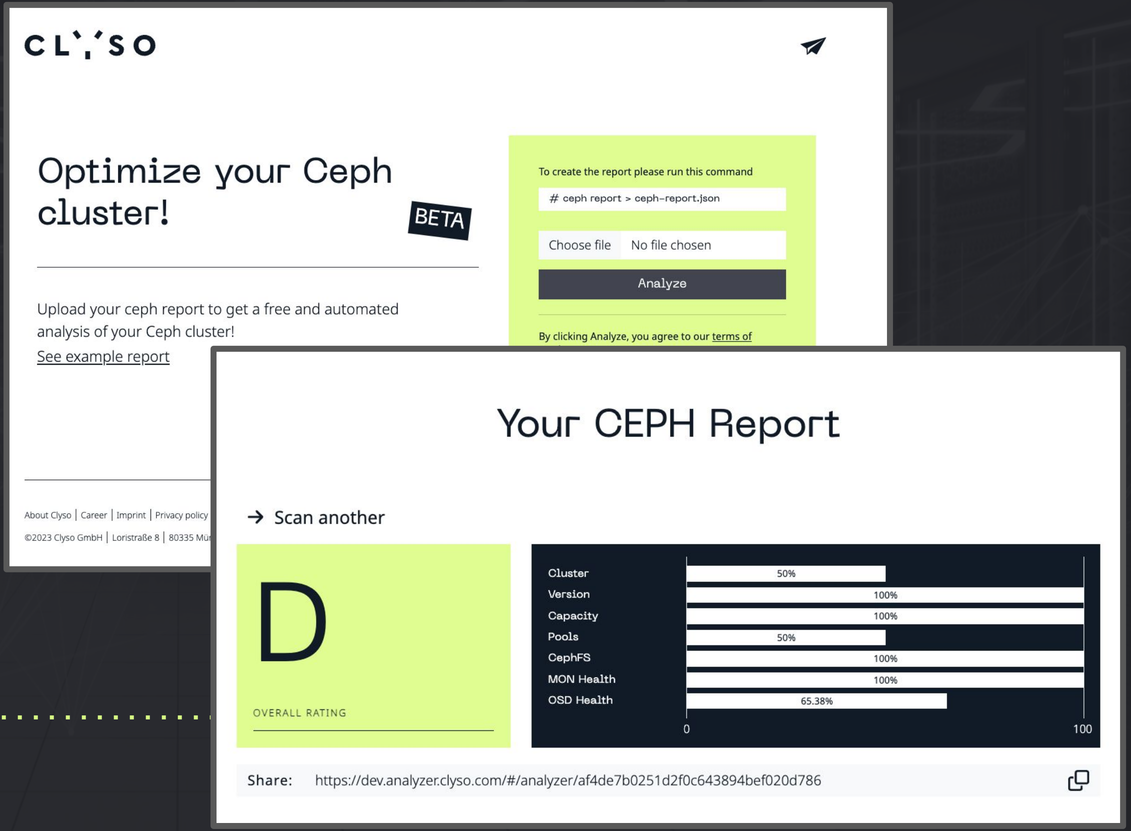


```
# Journal optimizations
# Note that these are for filestore. If you're using BlueStore, you should
filestore_min_sync_interval = 5  # adjust based on your hardware
filestore_queue_max_bytes = 10485760  # adjust based on your hardware
filestore_queue_max_ops = 5000
filestore_queue_committing_max_ops = 5000
filestore_queue_committing_max_bytes = 10485760  # adjust based on your har
filestore_wbthrottle_enable = false

# Make sure we're using the best possible allocator.
bluestore_block_size = 10737418240  # adjust based on your hardware
bluestore_bluefs_min = 5368709120  # adjust based on your hardware
bluestore_bluefs_max = 53687091200  # adjust based on your hardware
bluestore_cache_size = 10737418240  # adjust based on your hardware
bluestore_cache_kv_max = 1073741824  # adjust based on your hardware
bluestore_cache_kv_ratio = 0.3
bluestore_cache_autotune = true

# Optimizations for the MONs
[mon]
mon_osd_down_out_interval = 600
mon_osd_min_down_reporters = 5
mon_allow_pool_delete = true
```

Warning: do not use in production!

Ceph Cluster Analyzer

- I want to build tools that help people run Ceph.
- Step 1: a website which will grade your ceph cluster.
- Try it now:
  - https://analyzer.clyso.com
- Coming soon^{TM}
  - Clyso Enterprise Storage
  - Ceph Copilot
  - Chorus Multisite S3



Optimize your Ceph cluster!

BETA

Upload your ceph report to get a free and automated analysis of your Ceph cluster!

See example report

To create the report please run this command

```
# ceph report > ceph-report.json
```

By clicking Analyze, you agree to our terms of

About Clyso | Career | Imprint | Privacy policy
©2023 Clyso GmbH | Loristraße 8 | 80335 Mü

Your CEPH Report

Share: https://dev.analyzer.clyso.com/#/analyzer/af4de7b0251d2f0c643894bef020d786

Thank you

dan.vanderster@clyso.com