ceph days
LONDON 2026

From VMware, SAN, OpenAFS/NFS to Proxmox and Ceph/CephFS

A Practical Migration Path

ceph days LONDON 2026

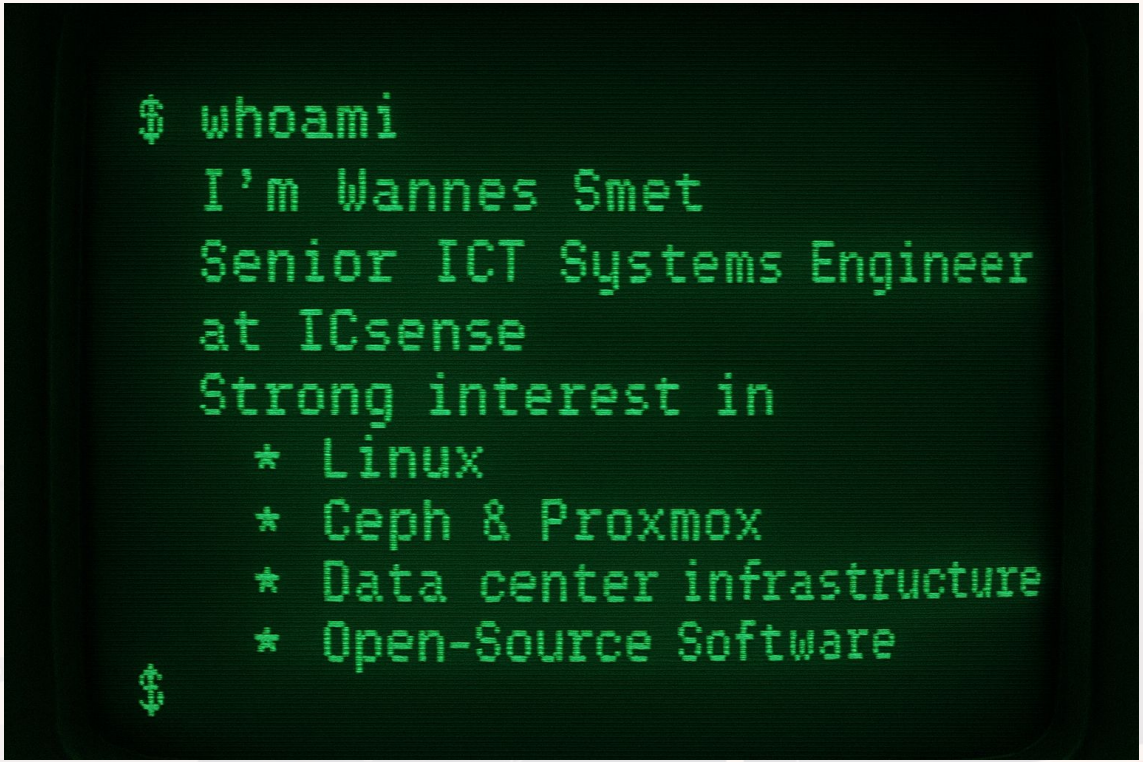
Agenda

- Introduction & context
- Starting point: VMware + SAN
- Why Consider Change
- Proxmox + Ceph
- Migration Path + Lessons learned
- Backup
- Any mishaps?
- Some final thoughts

Introduction & Context

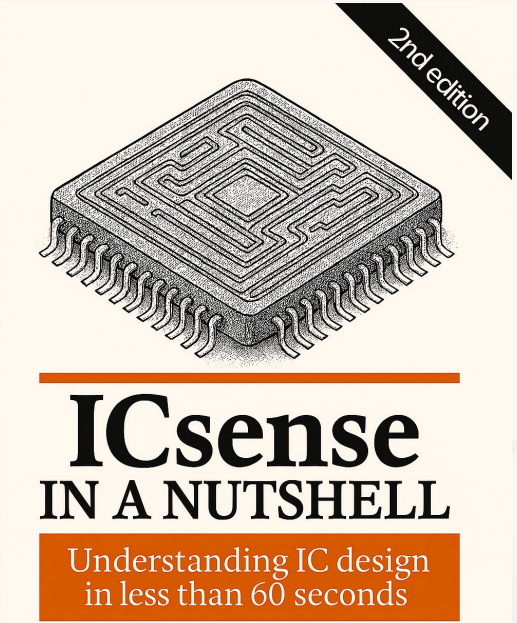
ceph days LONDON 2026

whoami



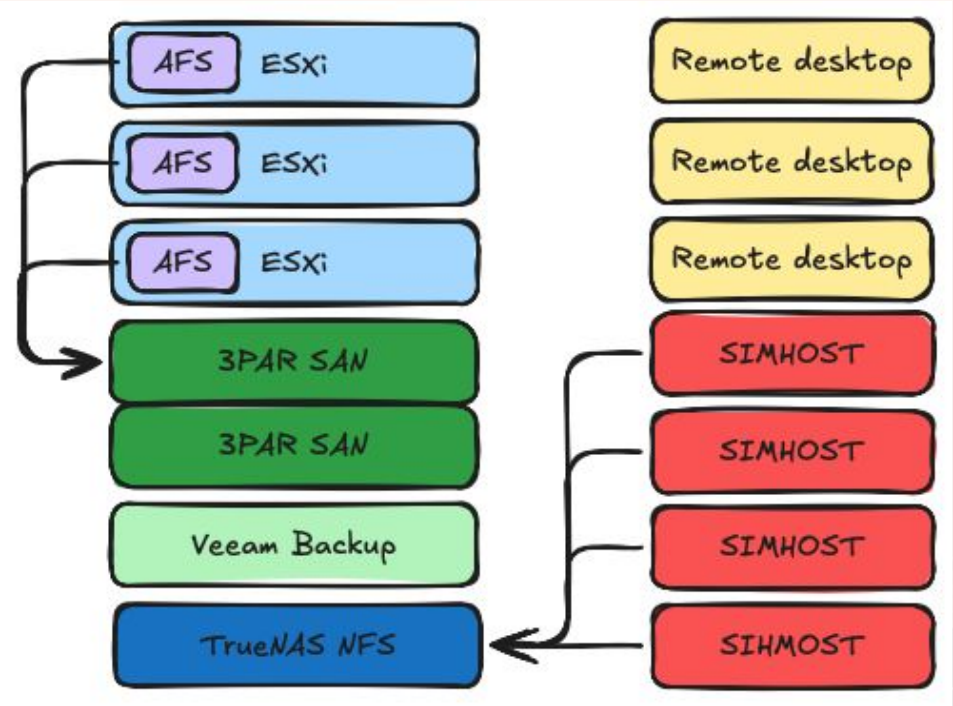
Introduction & Context

- IC / ASIC design company
- Extensive chip simulation for design verification
- Goal: eliminate design errors before fabrication
- Path from idea to tape-out is costly.
- ASICs don’t run code → no post-silicon fixes possible
- We're in medical, automotive, aerospace.
- Errors are not acceptable
- Fixes ≈ 6 months delay
- First-Time-Right and automation is key for us.



ceph days LONDON 2026

Starting point: VMware + SAN



- 3 ESXi servers, 512GB RAM, FC HBA, ~120VMs
- OpenAFS VMs for network file storage
- 3PAR SAN, 36 HDD 60TB net capacity (1200IOPS/20MBps throughput avg)
- Veeam Backup server (D2D2T)
- TrueNAS NFS for simulation data (gentle IO workload as well)
- ~10 Remote desktop machines
- ~20 simulation servers

Introduction & Context

How does that translate into “workload”?

- We want simulations to finish ASAP:
  - Many and fast cores (bare metal)
  - Lots of RAM
  - Lots of disk space
- Relatively frequent hardware refreshes for simulation workloads leave us with many still very capable decommissioned servers

ceph days LONDON 2026

Introduction & Context

How does that translate into storage workload?
7-day average measurement on our 3PAR SAN:

```
15:45:09 09/14/2023 r/w I/O per second    KBytes per sec    Svt ms   IOSz KB
         client:   24 MiB/s rd, 12 MiB/s wr, 3.04k op/s rd, 497 op/s wr    Domain
Cur Avg   Max   Cur   Avg   Max  Cur  Avg  Cur  Avg Qlen
                  -   t 1059 626  1081 17731 11545 64977 0.65 0.55 16.7 18.4    1
-------------------------------------------------------------------------------
                  1   t 1059 626       17731 11545       0.65 0.55 16.7 18.4    1
Press the enter key to stop...
```

For reference today:

```
 io:
    client:   24 MiB/s rd, 12 MiB/s wr, 3.04k op/s rd, 497 op/s wr
```

Why change the “winning” team?

ceph days LONDON 2026

2022: “VMware by Broadcom”

- Broadcom announced its intentions to acquire VMware
- We weren’t comfortable with that idea
- Renewed for 3 years at a ~20% price increase to buy us some time
- Look for alternatives.

ceph days LONDON 2026

An aging SAN

- Performance: more than sufficient
- 80% capacity
- Almost EOL: no upgrades/expansions possible
- Entry level SAN is ~affordable, but lacks features (Veeam integration)
- Mid range SAN: expensive and overkill

OpenAFS

- OpenAFS does scale, but not in a way that suits our workload.
- Linux, macOS and Windows all require custom kernel modules/extensions, drivers.
- Small install base, no “vivid community”
- Lots of problems on Linux and macOS:
  - Shell freezes (Kernel ring buffer: `waiting for busy volume`)
  - Memory leaks (resulting in crashing simulation servers)
  - macOS requires a kernel extension

ceph days LONDON 2026

NFS simulation data

File sharing NFS:

- One major flaw: SPOF

Enter Proxmox

ceph days LONDON 2026

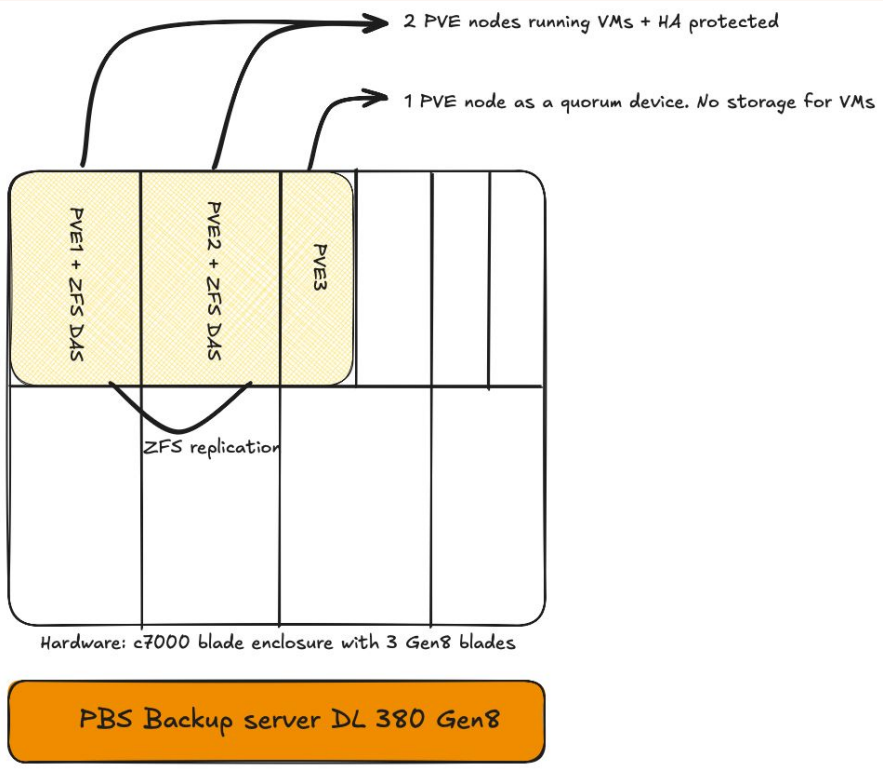
Enter Proxmox

Proxmox was high on our list from the start

- It fits our needs: run VMs, HA, backups, snapshots.
- Free □ You can test all you want.
- Runs on decommissioned simulation servers.
- Supports Ceph as an enterprise grade storage back-end 👍

ceph days LONDON 2026

2023: POC Proxmox cluster



Basic setup

- Decommissioned HPE c7000 blade enclosure
- BL460c gen8 blade (Xeon E5 Gen2)
- 2 nodes with replicated DAS ZFS pools
- 1 node as a “quorum device”, no storage
- HPE DL380 gen8 as a Proxmox Backup server (PBS)
- Ran some non-critical VMs.
- It just ... worked. Nothing to write home about during 2023-2024

Let’s talk Ceph!

ceph days LONDON 2026

Why Ceph?

Why Ceph is a good fit (or a better fit than ZFS)

- Ceph is Free and Open Source, no vendor lock-in, full control
- It runs on our decommissioned simulation servers as well, no HCL 👍
- Ceph= “true shared storage”, ZFS can offer “pseudo shared storage”.
- Ceph’s failure handling is elegant
- Ceph grows with our company.
- You can shape it to your needs.
- CephFS might also replace OpenAFS and NFS in our infrastructure

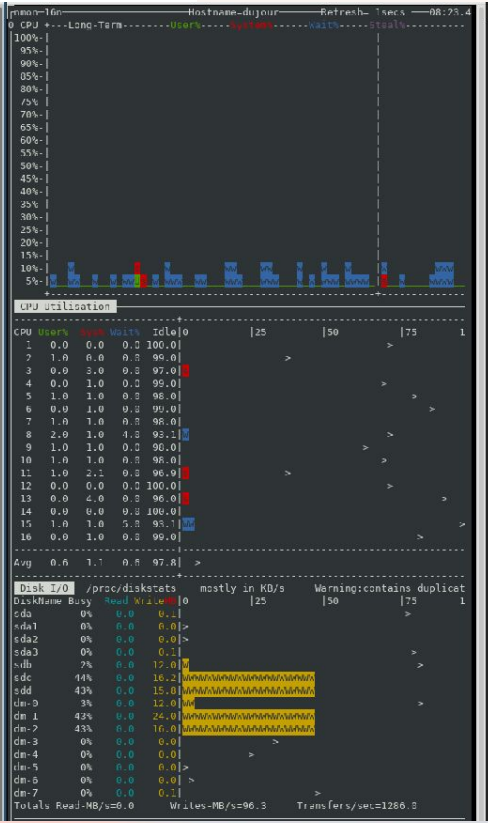
ceph days LONDON 2026

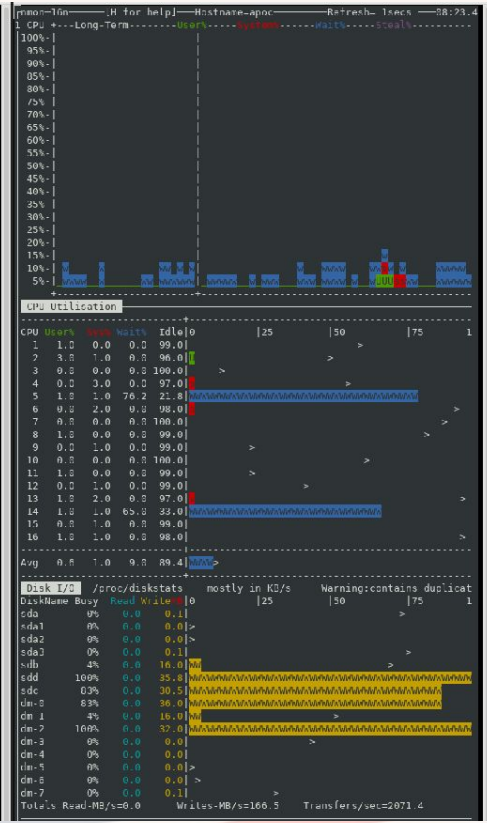
Jan 2025 – Enter Ceph

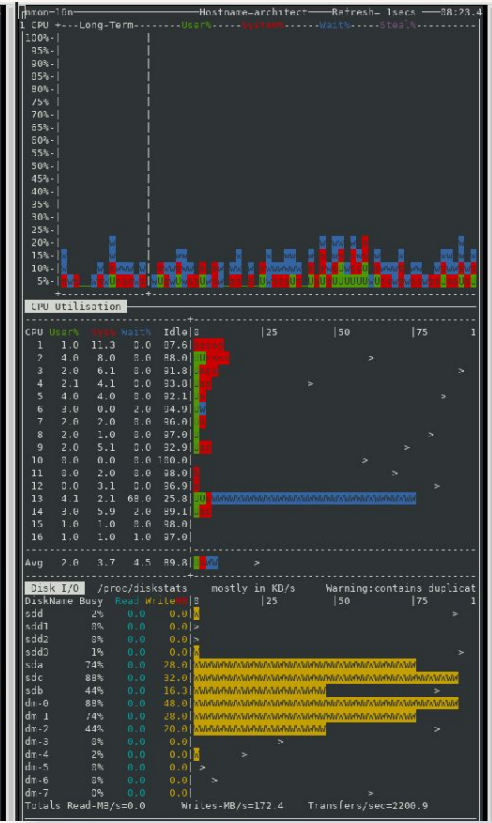
- I dove into the docs, so much going on, couldn’t wrap my head around “Ceph”.
- We invested in a 3-day Ceph training (highly recommended).
- The training really helped to make sense of the documentation.
- One insight gathered is to run an external Ceph cluster.
  - Avoid potential pitfalls of HCI (compute resource contention)
  - We want better knowledge of Ceph in case things go wrong.
- Time to bootstrap a POC Ceph cluster! 🎉

ceph days LONDON 2026

Ceph PoC stage 1





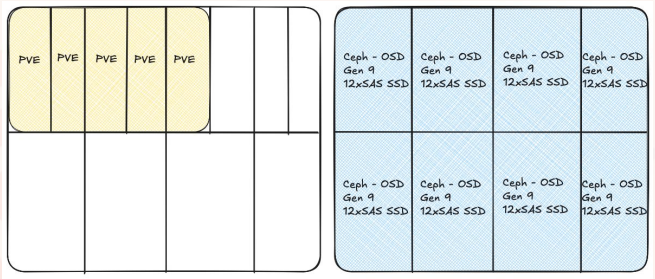


ceph days LONDON 2026

Q2/3 2025 - Ceph stage 2

I scale up and out.

- Carefully started migrating more low impact production VMs to the new setup.
- PVE from 3 to 5 nodes
- Removed the ZFS data stores in PVE
- 8 Ceph nodes, each 12 x 3.84TiB SAS SSDs
- 256GiB RAM, 16 cores@3.2GHz per node.
- 96 OSDs, 368TiB raw capacity

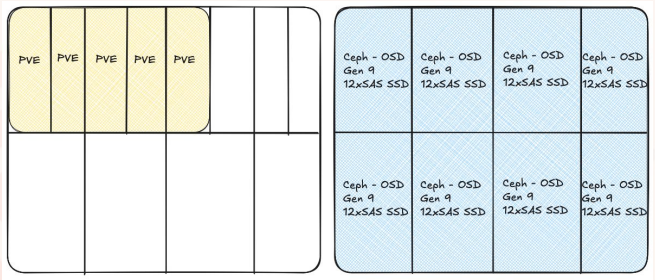


ceph days LONDON 2026

Q2/3 2025 - Ceph stage 2

Ceph has failure in mind. Our hardware/network stack had to match that:

- Upgrade from 10GbE to 20GbE
- Added a second NIC to each node, total of 4 “physical interfaces”
- 8 NICs, 2 Linux bonds for ceph-public and ceph-cluster
- CPU tuning got ping latencies from ~0.2ms to 0.04ms avg

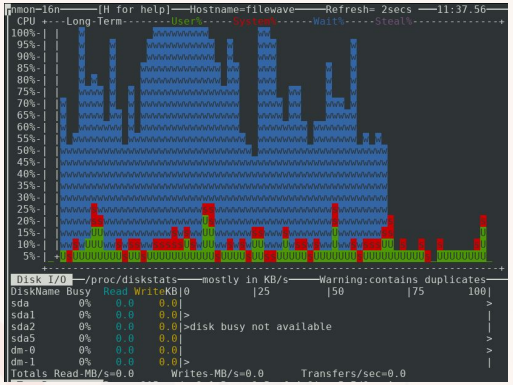


Main idea: I need many components to fail before production is affected.

ceph days LONDON 2026

Q2/3 2025 - Ceph stage 2

- PBS backups caused severe performance hits in VMs
- Workaround: reduce PBS workers.
- Root cause: Ceph public was routed over 1Gbit NIC
- Either change router or network topology of the cluster.



- Problem: changing the IP address of a mon is not straightforward
- In the end I decided to just destroy the cluster and bootstrap a new one.

ceph days LONDON 2026

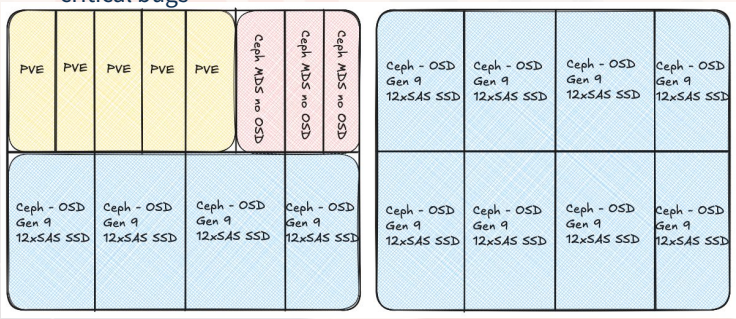
Q3 2025 - Ceph stage 3

File sharing OpenAFS:

- OpenAFS file share is a single large “directory”. My initial idea was to migrate to a similar large and monolithic CephFS filesystem.
- As my understanding of CephFS evolved, I shifted towards smaller, separate filesystems. (to contain blast radius)
- consequence: 2 pools/CephFS file system
- More pools > more PGs > more OSDs > need more block devices.
- “Intended” to add 48 extra SSDs
- 550TiB raw capacity
- Added 3 dedicated MDS nodes

Added a support contract

- During cluster intake we were made aware of the `bluestore_elastic_shared_blobs`bug in Ceph Squid
- Mitigated by redeploying all OSDs
- Nevertheless, postponed cluster expansion to X-mas maintenance.
- Main key take-away: If you start out with Ceph: Look up “known critical bugs”



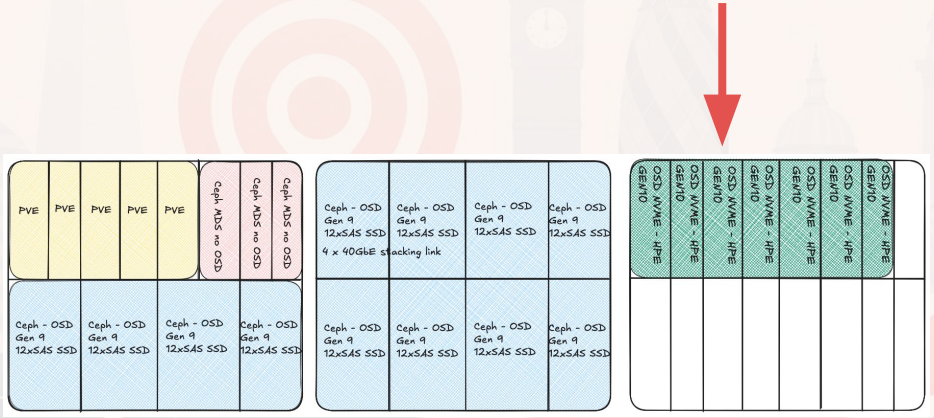
ceph days LONDON 2026

Q1 2026 - Ceph stage 4

Our dataset is very small-file heavy

- Irrelevant for backing up VMs (OpenAFS/CBT), but very relevant for CephFS.
- To keep backups windows within acceptable time frame, NVMe nodes/device-class was introduced for CephFS metadata pools.
- A third enclosure was added with 7 extra nodes.

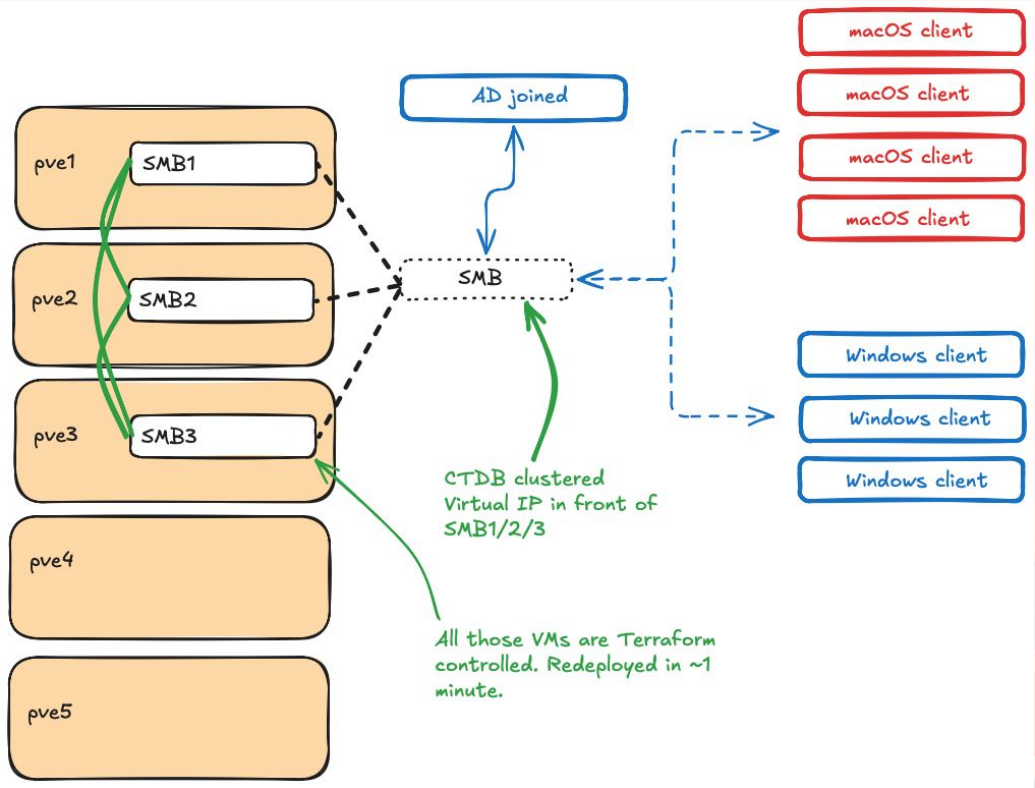
```
users@etxhost01: ~
$ cat /mnt/simulres/ | grep -e rbytes -e rfiles
 rfiles:                72084364
rbytes:           27538881819862
users@etxhost01: ~
$ echo "$((27538881819862/72084364/1024))kB per file"
373kB per file
users@etxhost01: ~
$
```



ceph days LONDON 2026

Q1 2026 – SMB setup

- macOS + Windows clients (native support)
- Samba **`vfs_fruit`** module is more than desirable for performance on macOS.
- I did not use the SMB manager module in Tentacle: it's brand new. We want more feedback from a larger install base.
- One EC-pool for archival data
- First user tests are promising!

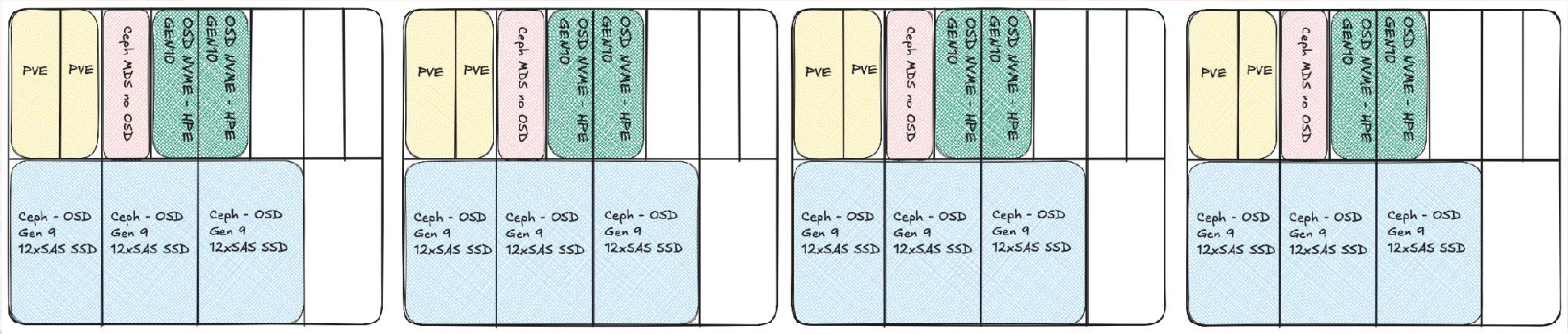


ceph days LONDON 2026

Q2 2026 - Ceph stage 5

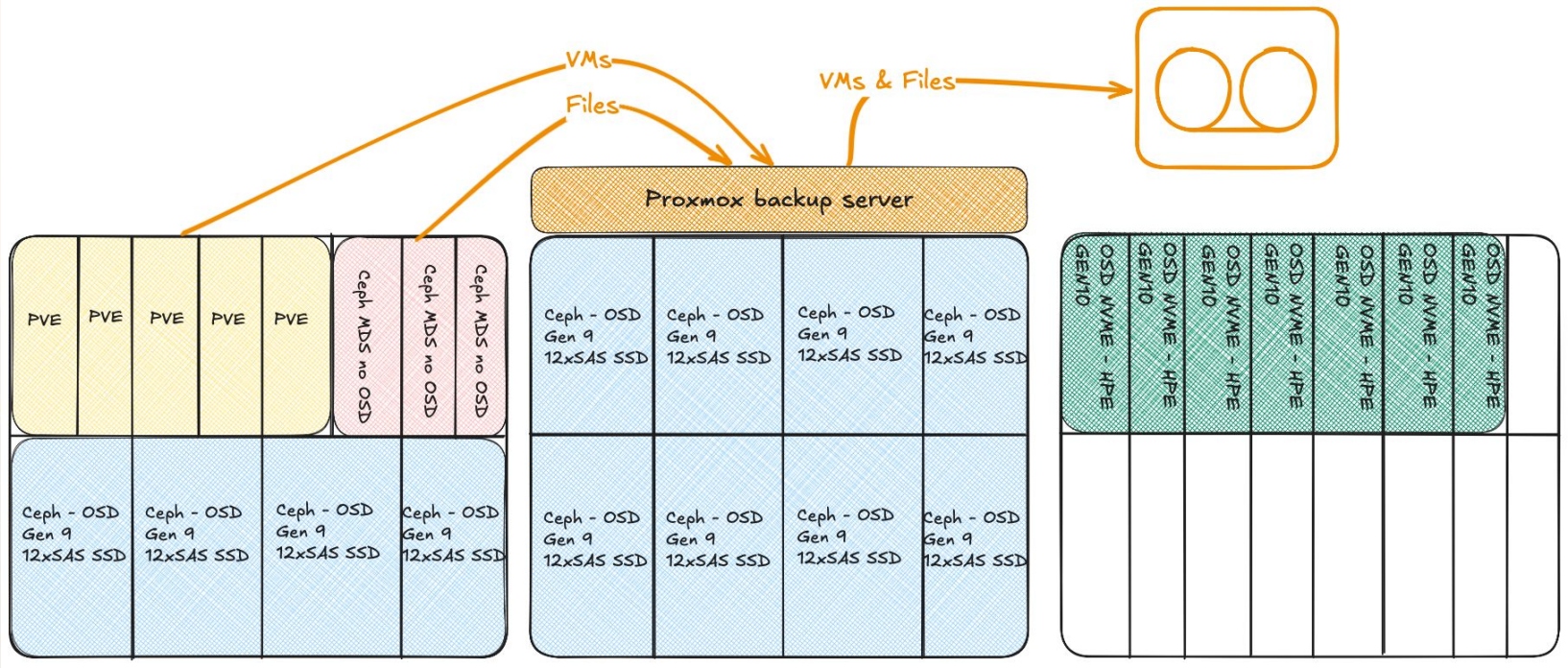
Scaling challenges: Failure domain & Layout

- Any enclosure crash is a recipe for disaster.
- To improve resilience, we want to add a 4th chassis, reshuffle hosts and bump the failure domain to “chassis”
- Challenges:
  - Injecting a new crushmap introduces a lot of IO



ceph days LONDON 2026

Backup?



Any Mishaps?

ceph days LONDON 2026

Any mishaps?

Some noteworthy mishaps

- `ceph orch host maintenance enter` does not check if there's enough available RAM on other nodes to restart an MDS elsewhere. Result: had to remount affected share on all clients.
- Uncontrolled cluster shutdown
  - Root cause: Power outage
  - Result: next slide

ceph days LONDON 2026

Any mishaps?

```
root@persephone:~# ceph -s
  cluster:
    id:     e9020818-2100-12a0-8aa2-9cdc71770100
    health: HEALTH_ERR
            unable to send alert email
            Invalid grafana certificate on host persephone: Invalid certificate key: [('PEM routines', '', 'no start line')]
            1 hosts fail cephadm check
            clock skew detected on mon.architect, mon.dujour, mon.seraph, mon.persephone
            2 osds(s) are not reachable

  services:
    mon: 5 daemons, quorum apoc,architect,dujour,seraph,persephone (age 4h)
    mgr: architect.nrvjah(active, since 4h), standbys: niobe.yhkgpe, seraph.guwwsy
    mds: 17/17 daemons up, 3 standby
    osd: 156 osds: 156 up (since 4h), 156 in (since 11d)
         flags noautoscale

  data:
    volumes: 11/11 healthy
    pools:   24 pools, 8225 pgs
    objects: 327.05M objects, 103 TiB
    usage:   280 TiB used, 240 TiB / 520 TiB avail
    pgs:     8224 active+clean
             1    active+clean+scrubbing+deep

  io:
    client:   0 B/s rd, 1.4 KiB/s wr, 0 op/s rd, 0 op/s wr
```

```
root@persephone:~# ceph health detail
…
[ERR] OSD_UNREACHABLE: 2 osds(s) are not reachable
    osd.53's public address is not in '192.168.11.0/24' subnet
    osd.86's public address is not in '192.168.11.0/24' subnet
```

ceph days LONDON 2026

Some final thoughts

What might make you think twice

- Ceph is not a set-and-forget
- The learning curve is real
- More in-house responsibility, rather than relying on your MSP.

Things we like!

- Disk space galore! 520TiB
- Future expansions are cost effective
- Ceph is flexible, we can shape/tune it to our needs
- We’re free from vendor lock-in!
- Investment shifts from hardware people, knowledge & skill.
- Active and helpful community: Slack, mailing lists, Reddit, Online AMA, … .