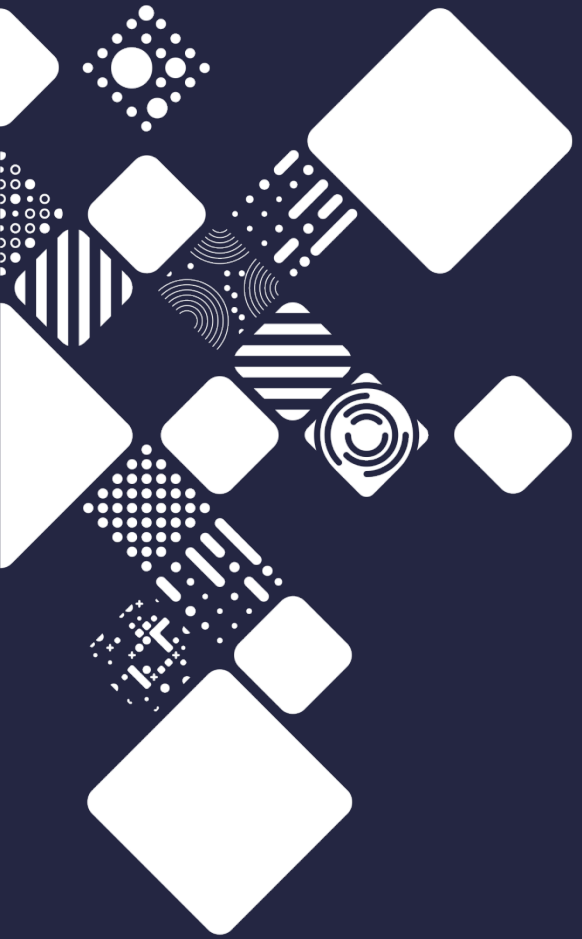


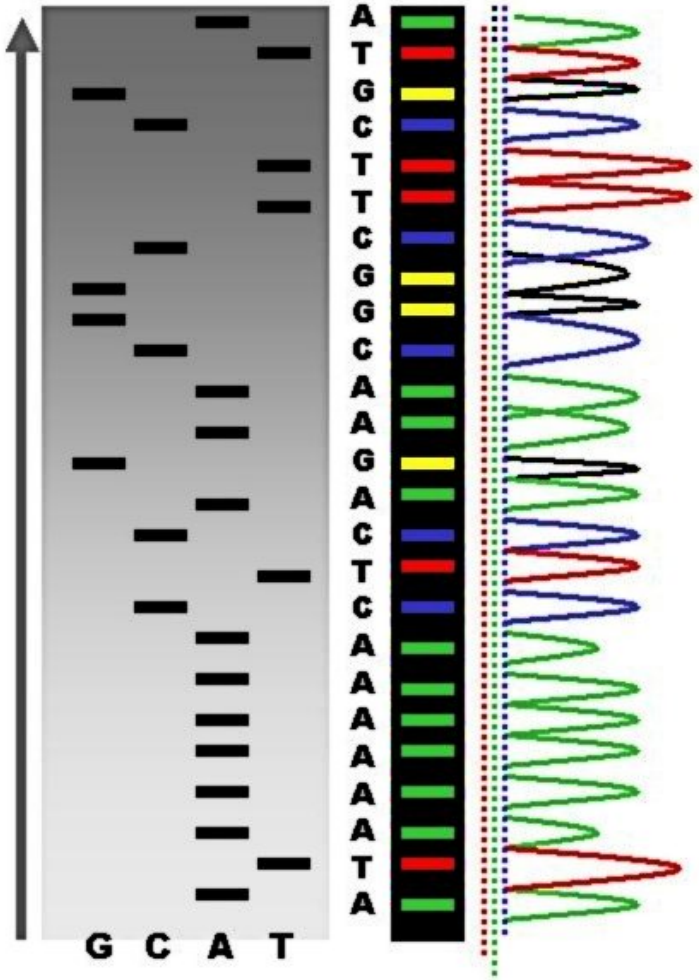
Dude, where's my cluster?

Dave Holland
dh3@sanger.ac.uk
Wellcome Sanger Institute

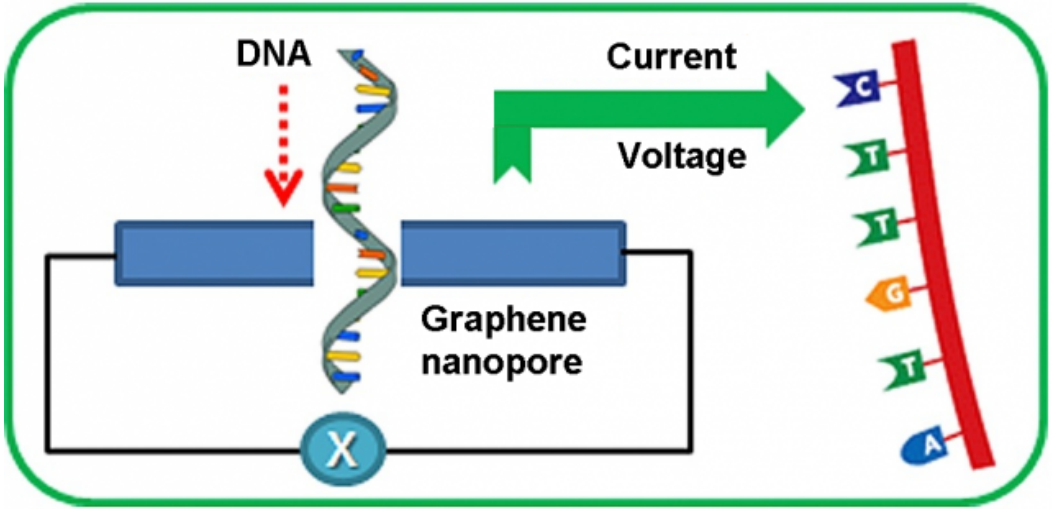
wellcome sanger institute

A long time ago in a ~~galaxy~~ datacentre far, far away....

wellcome sanger institute



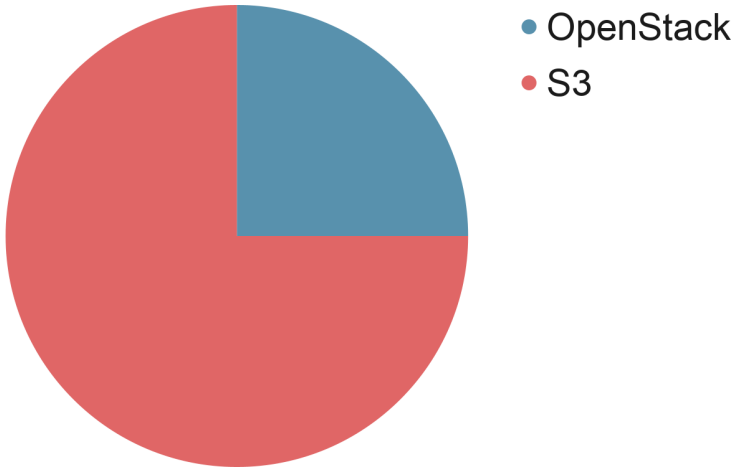
DNA sequencing (the 30-second summary)



wellcome sanger institute

Why Ceph

- We deployed Ceph for OpenStack.
- Users adopted it for S3.



wellcome sanger institute

In numbers

- ~5PB usable (15PB raw)
  - 51 OSD servers across 4 failure domains
    - 60 HDD OSDs per server = 3060 OSDs total
    - 8 NVMe OSDs for high I/O pools
- 6 Internet-facing radosgw servers using HAProxy for resilience
- Version 16 (Pacific)

wellcome sanger institute

wellcome sanger institute

- relocate some machines to free up “high power” cabinets
- replace the oldest 9 machines with new hardware
- keep the Ceph cluster up throughout

wellcome sanger institute

wellcome sanger institute

wellcome sanger institute

Power

- Ceph was originally installed in “high power” racks - 30kW burstable to 40kW
- They were the only racks available at the time due to partial fit-out
- A rack full of these Ceph servers “only” consumes 9kW
- Racks in other DC rooms can host up to 10kW
- Disproportionate cooling overhead compared to holding GPU servers (~30% excess overhead)

wellcome sanger institute

Lifecycle

- Replace the original 9 Ceph servers dating from 2017
- 6TB disks are “small” these days
- Increasing number of faults and maintenance costs

wellcome sanger institute

Availability

- Research storage
  ↓
- Dataset hosting
  ↓
- Collaboration
  ↓
- Mission-critical service

wellcome sanger institute

⚠ ceph-ansible

⚠ age of the hardware

⚠ networking changes

⚠ move duration

wellcome sanger institute

ceph-ansible

- Venerable a.k.a. abandoned - last release in 2022, “wontfix” everywhere, wants specific Python library versions
- Complicated (not helped by our own choices, e.g. dynamic inventory)
- It’s slow

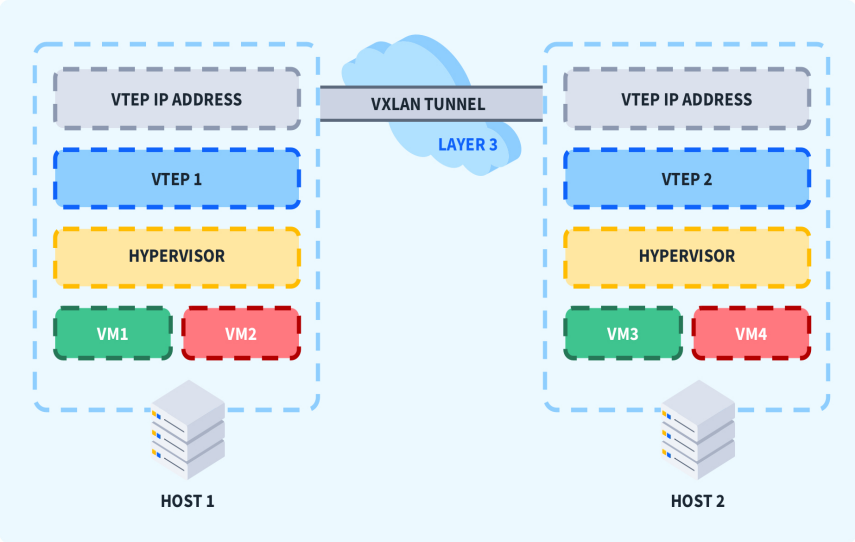
Age of the hardware

- The oldest machines were installed in 2017
- Increased disk failure rate during and shortly after the moves
  - approximately 2-3x higher
- Unexpected CPU and memory failures

wellcome sanger institute

Networking

- Fortunately we already had VXLAN for OpenStack
- “Stretching” the VLANs possible without being ugly
- **Same VLAN, same IP addresses, same hostnames**
- Transitional names for IPMI interfaces



wellcome sanger institute

Move duration

- Low-priority project, long overall duration - September 2025 to January 2026
- Occasional fits of “where did we get to?”
- Christmas change freeze got in the way

wellcome sanger institute

- moves
- rebuilds
- multipath
- missing NVMe
- deep scrub schedule

wellcome sanger institute

Moves

- Set noout and norebalance and “**Just Do It**”
- Occasional confusion over 1G vs IPMI connection
- Physical disturbance
  - nearby hypervisors were not correctly latched into the rack
  - power leads without latches can work loose

Rebuilds

- We took the short cut: shut down each old machine, installed its replacement with ceph-ansible, and let the rebuild run
- Experimented with recovery throttles

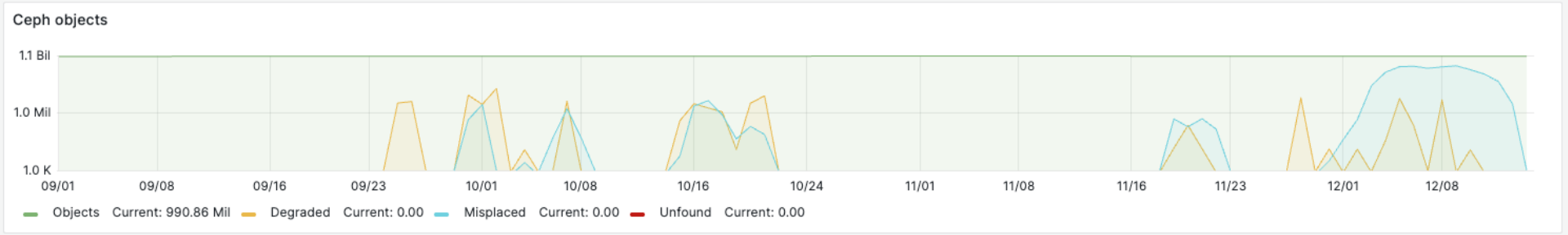
```
ceph tell osd.\* injectargs --osd_max_backfills=3 --osd_recovery_max_active=12
```

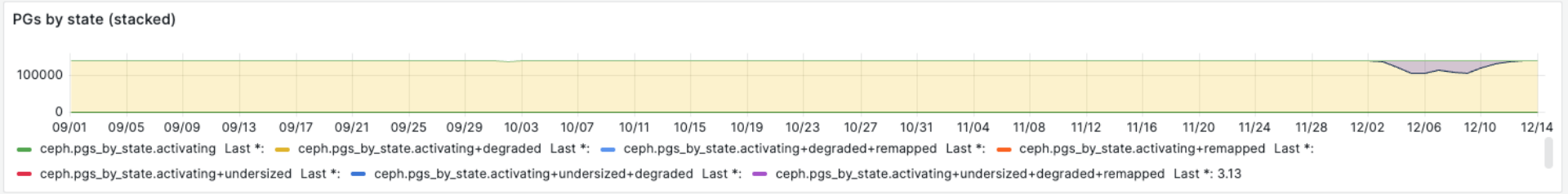
wellcome sanger institute

Rebuilds

- Although throttled, the high rebuild traffic might have contributed to disk failures on older machines
- Bigger HDDS in new machines (18TB vs 6TB) increased cluster raw capacity from 15PB to 23PB

Rebuilds





wellcome sanger institute

Rebuilds

- We probably did too much too quickly occasionally…

```
root@sto-1-1:~# df -h /var/lib/ceph/mon/
Filesystem                                Size  Used Avail Use% Mounted on
/dev/mapper/sto--1--1--nvme-cephmondir  1.8T  319G  1.4T  19% /var/lib/ceph/mon
```

wellcome sanger institute

Multipath

- New SCSI controller (“no, you don’t really have 120 HDDs”)
- ceph-ansible needed patching
- Our local OSD management scripts needed updating

```
mpathz (35000c500f43cdbeb) dm-28 SEAGATE,ST20000NM002D
size=18T features='0' hwhandler='0' wp=rw
|-+- policy='service-time 0' prio=1 status=active
| `- 0:0:5:0   sdf     8:80   active ready running
`-+- policy='service-time 0' prio=1 status=enabled
  `- 0:0:36:0  sdaj    66:48  active ready running
```

wellcome sanger institute

Missing NVMe

- Some new machines were accidentally ordered with only 2x NVMe - only sufficient for OSD WAL partitions
- But we also needed NVMe for RGW metadata pool OSDs (among others)
- Scavenge and recycle!

Deep scrub schedule

- Interval was 28 days due to large number of OSDs/PGs
- We had set nodeep-scrub to speed up recovery
- Thousands of PGs got delayed beyond 28 days
- Manual deep scrub initiation over a couple of weeks to regain the spread

```
root@sto-1-1:~# ceph pg dump | grep active |
awk '{print $23}' | sed 's/T.*//' | sort | uniq
-c | sort -k 2 | head -10
dumped all
   1098 2025-12-01
   1329 2025-12-02
    976 2025-12-03
    445 2025-12-04
     14 2025-12-05
     63 2025-12-06
    144 2025-12-07
     96 2025-12-08
     52 2025-12-09
    587 2025-12-10
```

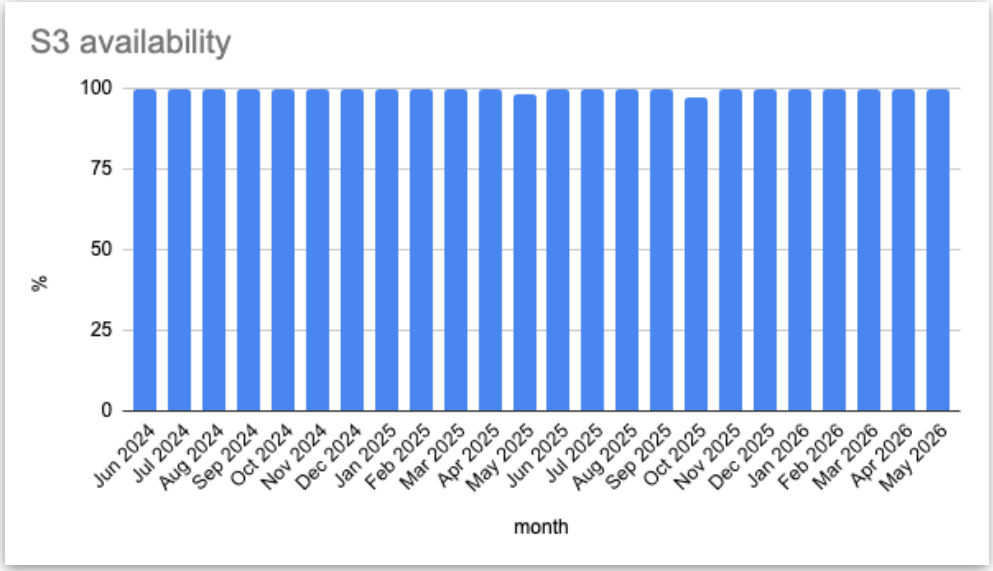
- relocate some machines out of the “high power” cabinets ✅
- replace the oldest 9 machines with new hardware ✅
- keep the Ceph cluster up throughout ✅

wellcome sanger institute

In numbers

- Ceph is generally hugely robust!

Sep 2025	100	
Oct 2025	97.251	Failing/slow disk
Nov 2025	100	
Dec 2025	99.989	
Jan 2026	99.98	
Feb 2026	100	



wellcome sanger institute

- get off ceph-ansible
- upgrade, upgrade, upgrade
- more hardware lifecycle replacements (or retirements)

wellcome sanger institute

Thanks for listening

- Happy to take questions now or by email: dh3@sanger.ac.uk

wellcome sanger institute