Elettra Sincrotrone Trieste
Life *(of big storage)* in the fast lane
“Data expands to fill the space available for storage”

Computed Tomography
*Higher*
- needed quality
- frequency scans
- resolution

**4DCT**
- 10TB/day Elettra
- 100TB/day Elettra2.0

1PB/year easily

**Sensors**
- from kilo to megapixels
- 5Hz ... 120Hz ... kHz!
From Storage Big Bang...

- DEC AlphaServer 2100
- EMC2 CX4-240
- IBM’s GPFS
- GLUSTERFS
- BigBench (GLUSTER)
- Sofa (CEPH)
… to a GLUSTER Big Bench

**PROS**
- ✔ Lightweight and simple
- ✔ High throughput with big files
- ✔ No metadata

**CONS**
- ✔ Debugging and Recovery
- ✔ Poor performances
- ✔ No metadata
“OSDs may see a significant performance improvement by storing an OSD’s journal on an SSD and the OSD’s object data on a separate hard disk drive.”

- 20 x 2.5” 600GB 15kRPM
- 4 x 800GB HGST NVMe
  (1 SSD serves 5 OSDs)

- 20 x 6TB 7200RPM
- 4 SSD for journaling
  (1 SSD serves 5 OSDs)
SSD/NVMe drawbacks
✓ TBW limit of SSD’s ... 1yr lifespan
✓ 1 SSD failure affects 5 OSDs
✓ SSD defective stock disaster
✓ Replica 2 is not enough

Kernel bugs (4.8.10 vanilla)
✗ NMI watchdog: BUG: soft lockup - CPU#14 stuck for 23s! [kswapd1:157]
✗ Upgraded. Vanilla vs CentOS standard?

Domino effect
➢ common/HeartbeatMap.cc: 79: FAILED assert(0 == "hit suicide timeout")
➢ Stubborn processes fixed on dead OSD, ignoring its replicas
➢ Hammer bugs?

Can see a Luminous light
➔ 8 new nodes, replica 3, ~ 1PB net
➔ get rid of journaling SSD’s
➔ stable Bluestore improvements
Thank you!

www.elettra.eu