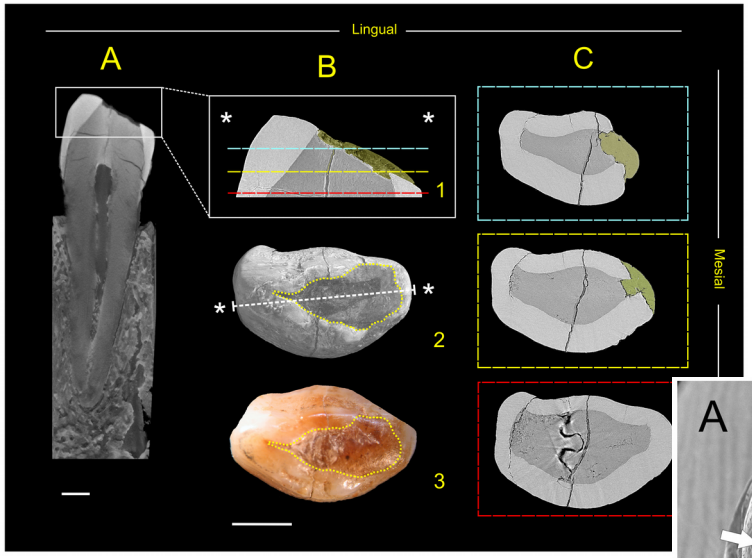
Elettra Sincrotrone Trieste

Elettra
Sincrotrone
Trieste

Life (*of big storage*) in the fast lane

Elettra Sincrotrone Trieste

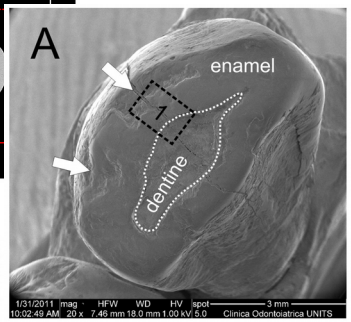
"Data expands to fill the space available for storage"

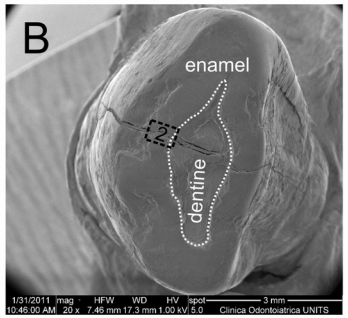


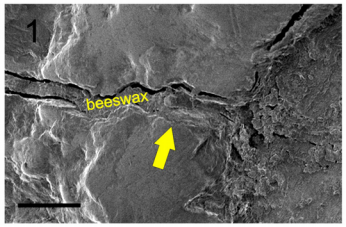
Computed Tomography

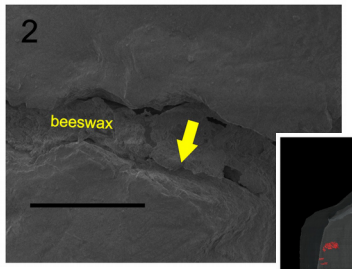
Higher

- ✔needed quality
- ✔frequency scans
- ✔resolution





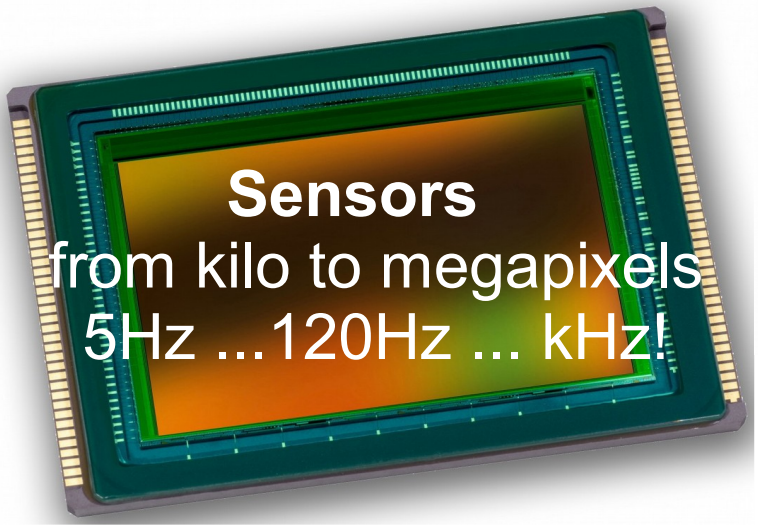




4DCT

- ✔10TB/day Elettra
- ✔100TB/day Elettra2.0

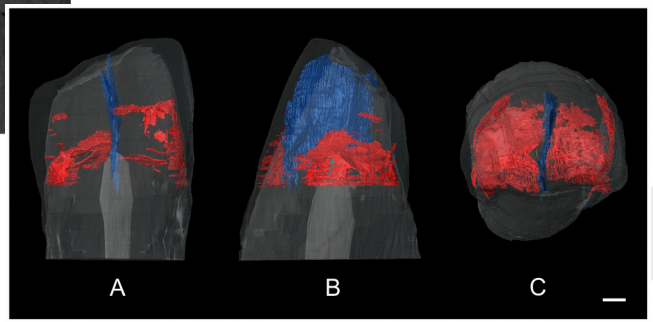
1PB/year easily



Sensors

from kilo to megapixels

5Hz ...120Hz ... kHz!



Elettra
Sincrotrone
Trieste

From Storage Big Bang...

IBM's GPFS

DEC AlphaServer 2100

EMC2 CX4-240

Sofa (CEPH)

1995 **2000** 2005 2010 2015 2020

Several PC's

BigBench (GLUSTER)

GLUSTERFS

Elettra Sincrotrone Trieste

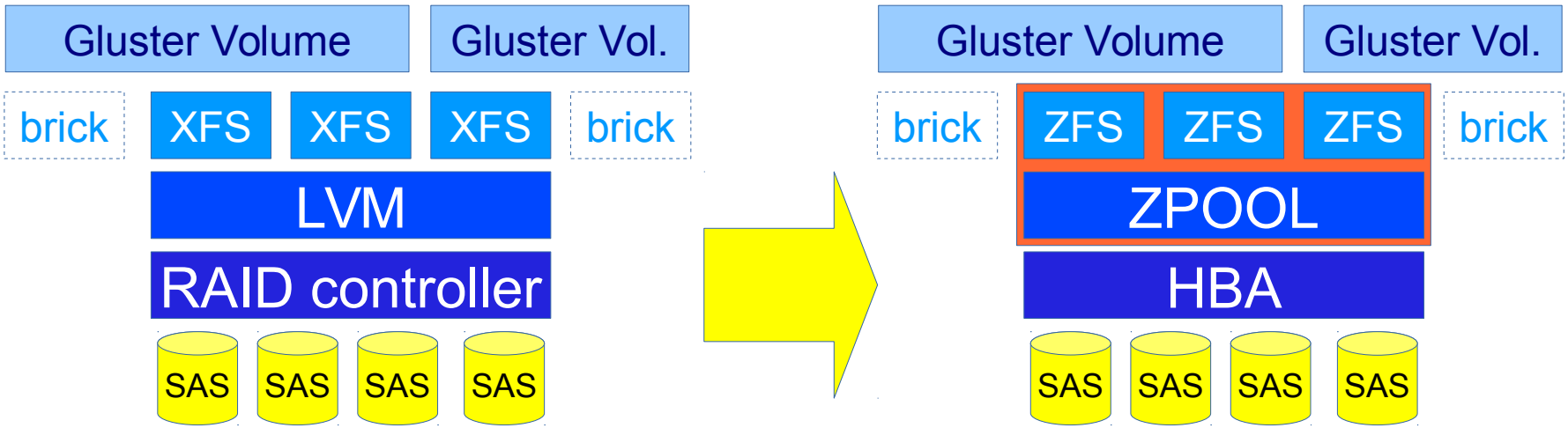
… to a GLUSTER Big Bench

PROS

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

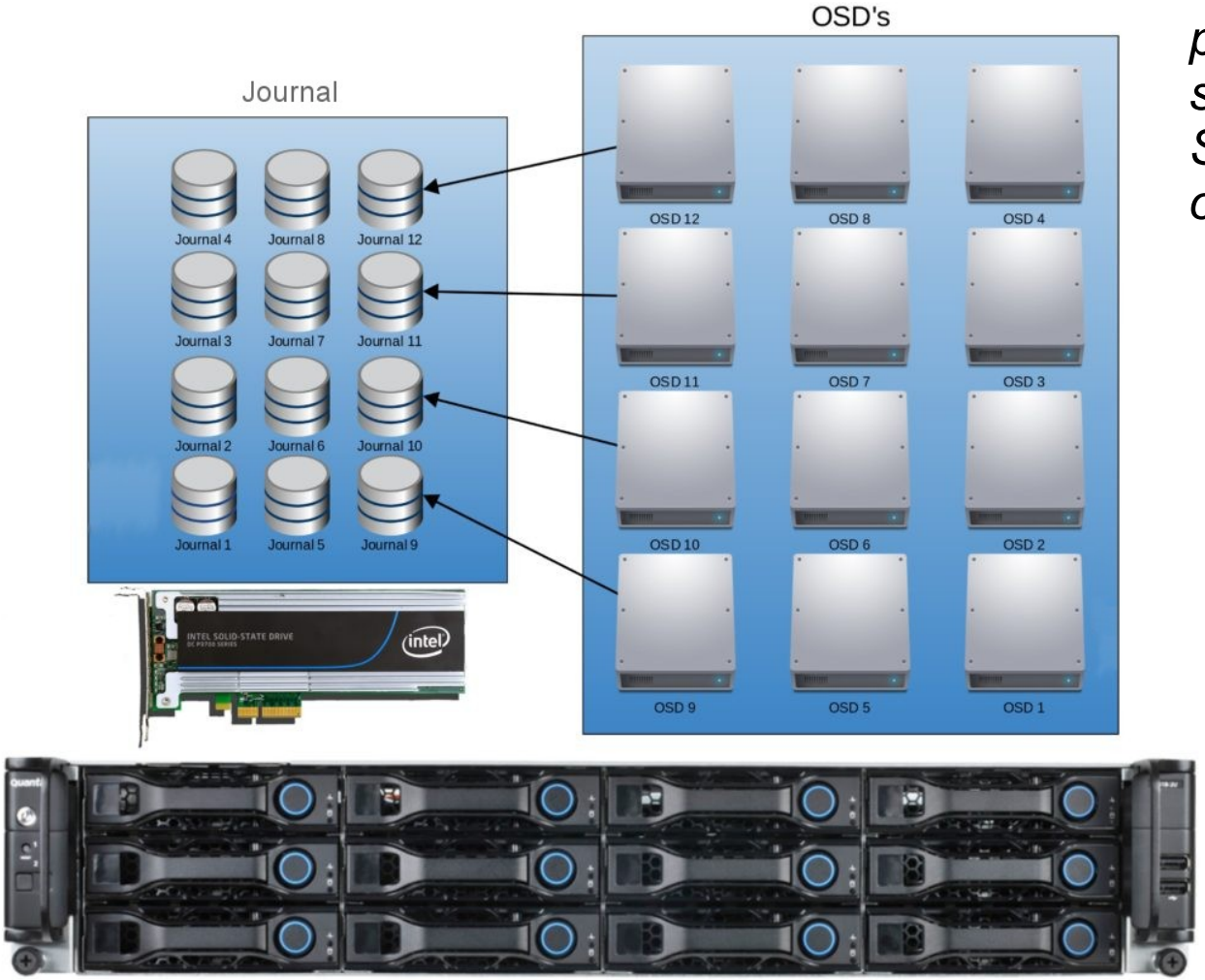
CONS

- ✔ Debugging and Recovery
- ✔ Poor performances
- ✔ No metadata



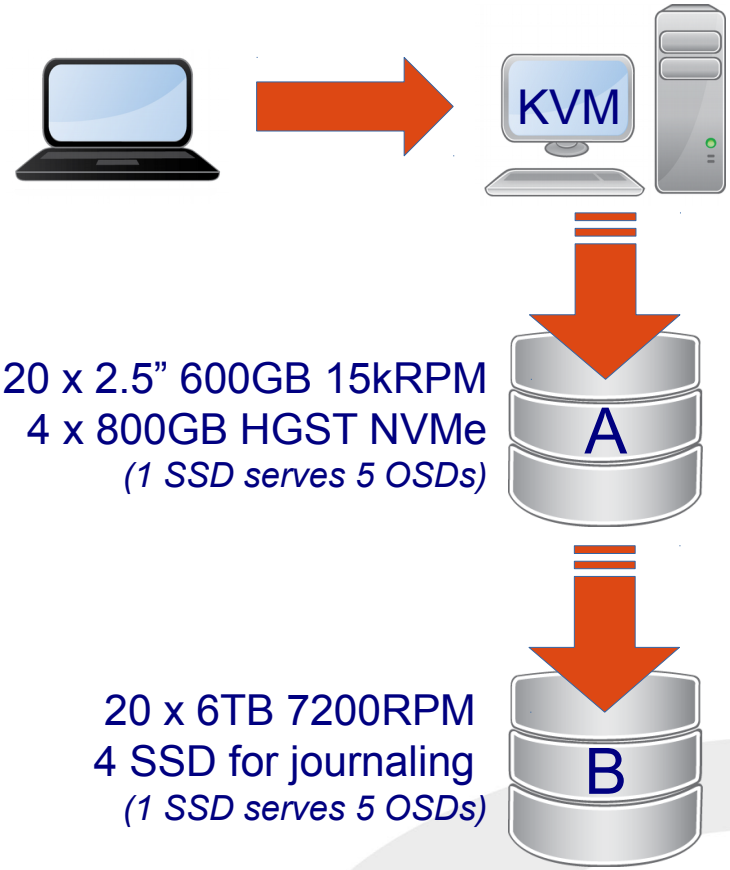
Elettra Sincrotrone Trieste

Laying on a Sofa!



ceph

"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)

Elettra Sincrotrone Trieste

Anything that can go wrong...

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

Elettra
Sincrotrone
Trieste

www.elettra.eu