DIGITAL PRESERVATION WITH OPEN SOURCE

SAGE WEIL
CEPH PRINCIPAL ARCHITECT
RED HAT

WHAT IS CEPH

- Scale-out distributed storage

  - terabytes to exabytes

- Self manage whenever possible

- Fault tolerant – no single points of failure

- Storage hardware agnostic

- Commodity components

- Single cluster, multiple protocols

  - Object, Block, File

- Free and open source

VALUE OF OPEN SOURCE FOR ARCHIVAL STORAGE

- Cost at scale

- Hardware vendor independence

  – Drives down cost

  – Price vs performance vs robustness

- Software vendor independence

  – Data lifetime far exceeds vendor lifetime

- Transparency

  – How do you read your data in 10, 20, 50 years?

  – Data is not hostage to proprietary platform

- Efficient investment of tax dollars

  – Technology investment benefits all users, not a single vendor

ARCHITECTURE TRENDS

- Hybrid SSD (flash) and HDD (hard disk) architectures

  - Range of flash solutions

    - Capacity, write, or read optimized

  - Tiering

- Ethernet attached devices

  - HDDs and SSDs attached directly to network

  - Embedded low-power processors running Linux

  - Eliminate "host" systems with expensive x86 CPUs, RAM

  - Several products announced, many others coming
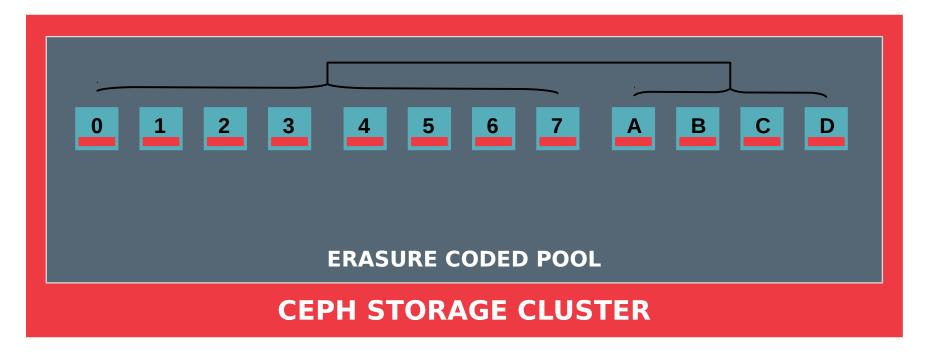
ERASURE CODING

- Ceph recently added erasure coding support

  - Pluggable algorithms

  - Other open storage systems are doing the same

    - Swift, GlusterFS

- Trade-off between space efficiency and more expensive recovery

  - Can be problematic for low-power, low-cost devices

- Local Recovery Codes

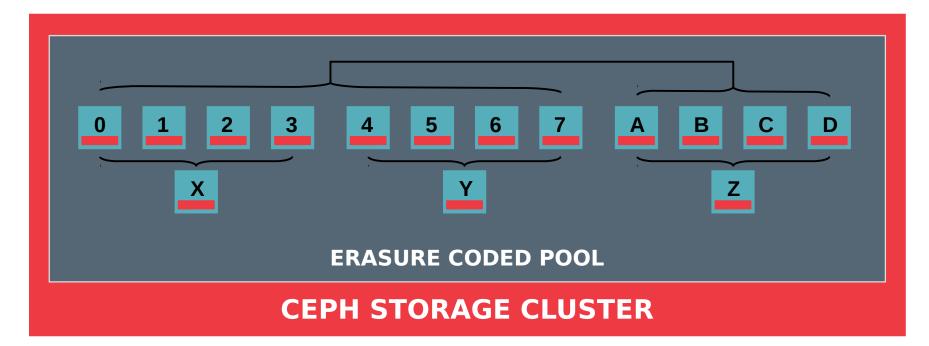 - Trade some extra storage for recovery efficiency

ERASURE CODING

OBJECT

ERASURE CODED POOL

CEPH STORAGE CLUSTER

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | A | B | C | D |

Recover with any 8 of 12

1.5x

LOCAL RECOVERY CODE (LRC)

OBJECT

0 **1** **2** **3** **4** **5** **6** **7** **A** **B** **C** **D**

X **Y** **Z**

ERASURE CODED POOL

CEPH STORAGE CLUSTER

Recover any one block with 4 of 5
Otherwise, recover with any 8 of 12

1.87x

OPTIMIZING FOR POWER

- Electricity has become a cost driver

- Offline storage attractive

  - One advantage of tape and optical disc media

  - Slow access times (minutes)

- Online devices (SSD, HDD) can be powered down

  - Not as cheap, but faster access times (seconds)

- Not yet present in open SDS projects, but coming

  - Erasure codes + low-power ethernet attached devices + adaptive power utilization

THANK YOU!

Sage Weil
CEPH PRINCIPAL
ARCHITECT

✉ sage@redhat.com

🐦 @liewegas

ceph