EOS Open Storage

EOS as a tape storage replacement

EOS workshop

4-5 February 2019
CERN
Europe/Zurich timezone
There is a live webcast for this event.

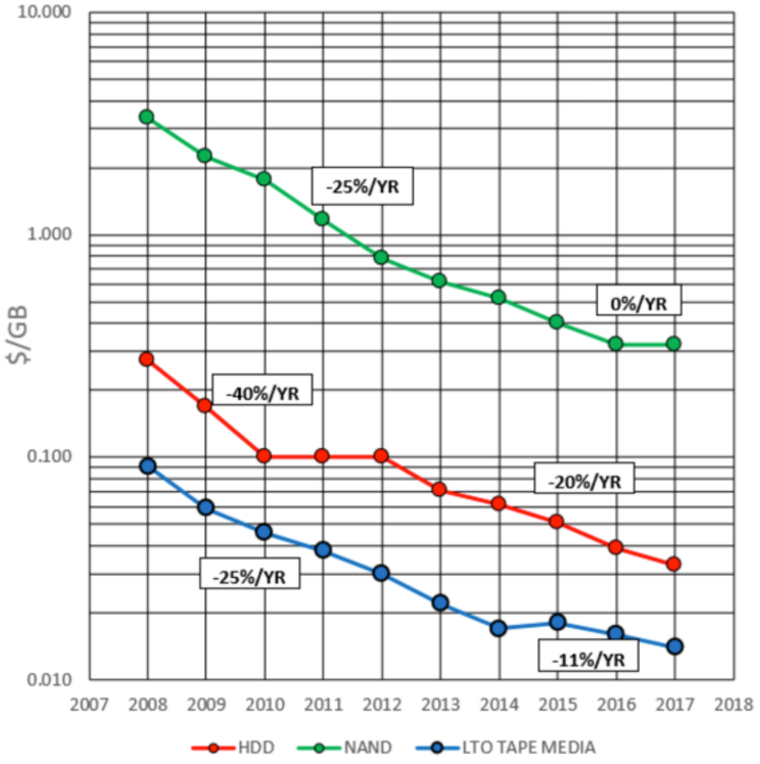
http://eos.cern.ch

Andreas-Joachim Peters
CERN IT-ST

An old story . . .

	Disk (HHD/SSD)	LTO (Tape)	Cloud
Initial Investment	Moderate	High	Low, but constant expense
Maintenance Cost	High	Moderate	Included
Expandability	Easy to Moderate depending on workflow and set-up	Easy to add additional tape media	Easy, but incurs higher monthly fees
Access Time	Fast	Moderate (with loading time)	Slow (depends on connection)
Sensitivity	High sensitivity, relatively high failure rate, maintenance required for off-site and long-term storage	Low sensitivity, low failure rate, low maintenance for off-site and long-term storage	Low sensitivity, low failure rate, low maintenance for off-site and long-term storage
Ideal Uses	Good for backup and failover, long-term storage for moderate amounts of data	Good for long-term archiving of large amounts of date	Good for archiving and backup if quick retreval is not required, cost prohibitive for large amounts of data

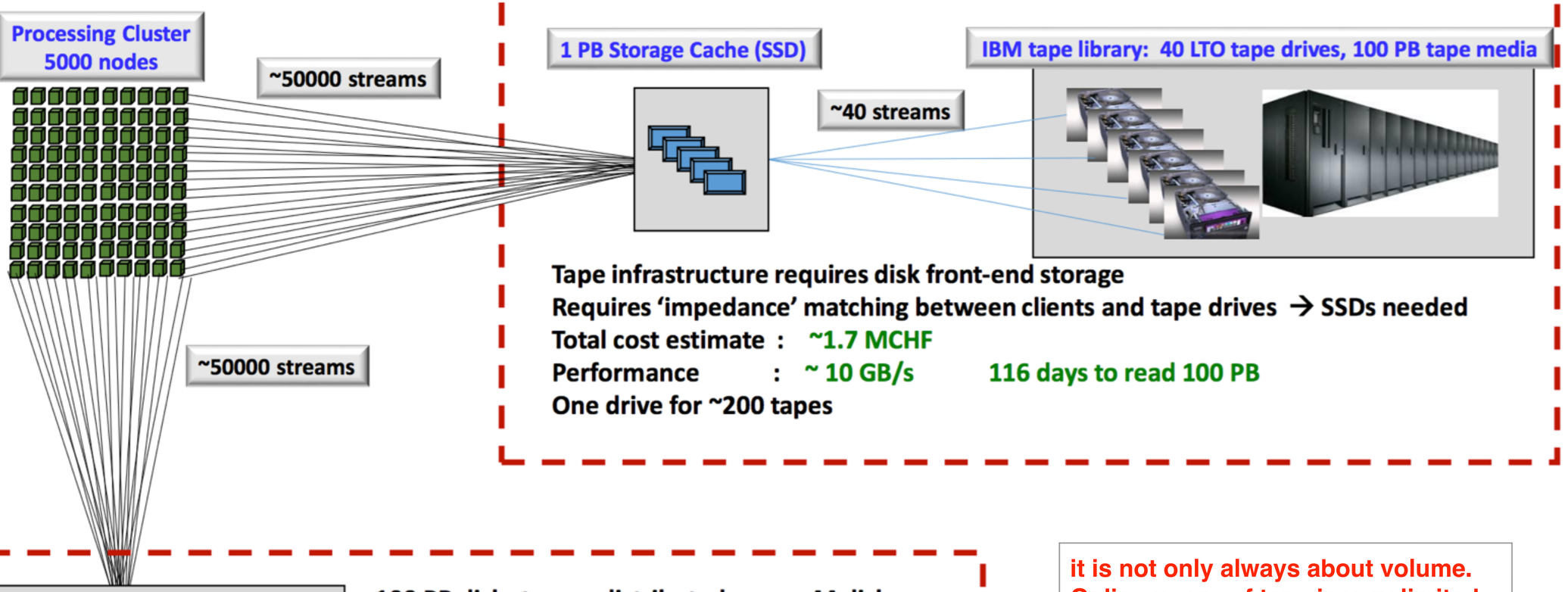
disk vs. tape vs. cloud



Cost Comparison

Bernd Panzer Steindl - CERN IT

100 PB storage example I



Tape infrastructure requires disk front-end storage
Requires 'impedance' matching between clients and tape drives → SSDs needed
Total cost estimate : ~1.7 MCHF
Performance : ~ 10 GB/s 116 days to read 100 PB
One drive for ~200 tapes

100 PB disk storage distributed across 44 disk server
Total cost estimate : ~3.5 MCHF
Performance : ~440 GB/s
3 days to read 100 PB
One server for ~200 disks

it is not only always about volume. Online usage of tape is very limited, while a disk based system can serve online & offline purpose!

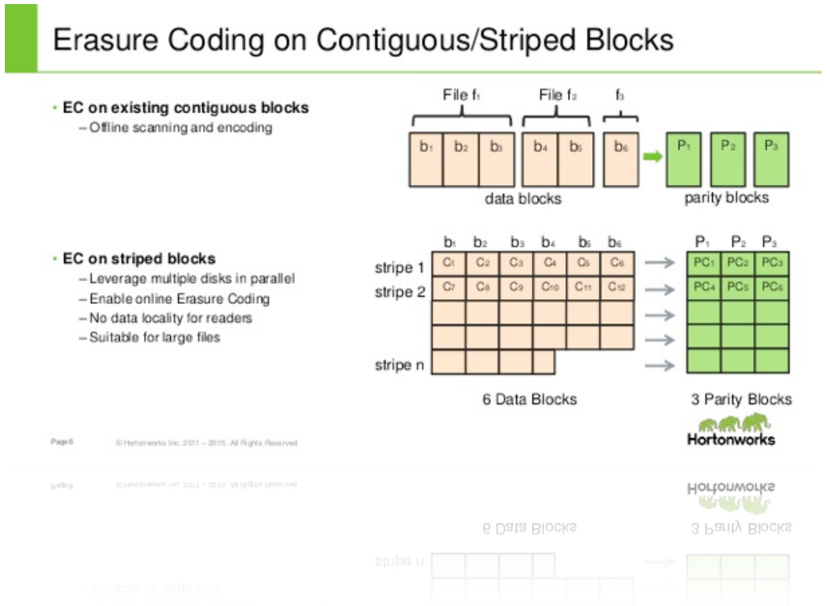
Future of tape market with single vendor became less predictable/reliable

Disk as Tape

- **envisaged characteristics**
  - **protect against dataloss** through ‘expected’ disk failures
  - design system with **minimal loss/impact** in case of a fatal failure
  - provide a flexible configuration to **scale availability against price** economy
  - optimize for **GB/$**
  - optimize for **large file streaming** access

Erasure Coding

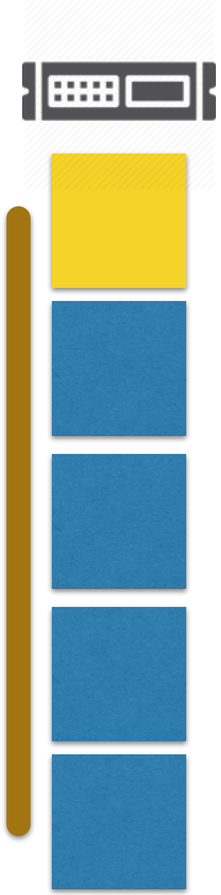
Models



- The **cheapest** erasure coding mode is **RS(M,K)** where M >> K e.g.
  RS(32,3) : ~10% cost overhead allowing 3 parallel failures without data loss
  - efficient only for streaming writes of large files (>128 MB)
  - would operate only in a **no repair** and **stream-only** mode because repair traffic too high (32 x single disk size)
    - reading will require **CPU** for reconstruction at first disk failure (RS 5GB/s/core)
  - EOS currently limits to **M+K <=16** (can be changed in software) because intention is to make redundancy over disks on different nodes, not on disks within a node
  - Alternative to EOS-RAIN (RS):
    ZFS RAIDZ-3 would work if redundancy is done within one node (disk recovery would require 1day @ 4.4 GB/s read speed - hypothetical)

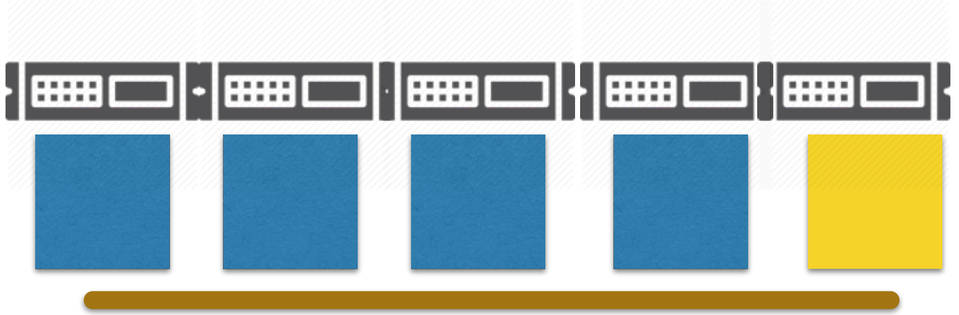
Group as many disks together for price economy - but …

Erasure Coding Models



ZFS RAIDZ
RAID6
EOS-RAIN

vertical EC



EOS-RAIN

can use either
- **horizontal**
- **vertical**
- **2D erasure encoding**

horizontal EC

Erasure Coding Models

10 nodes x 192 disks	vertical	horizontal	2D		Capacity
(29,3)	90.6%			unavailable on node failure	20.8 PB
(6,3)		66%		resilient against 3 disk and node failures	15.2 PB
h(29,3) + v(8,1)			80.6%	extremely low probability for unavailability or data loss	18.5 PB

Storage capacity [%] compared to RAW capacity 23 PB

Cost of highly redundant storage setups

Possible EOS Optimisations as Cold Storage

- Make unwanted **deletion extremely difficult**
  - add file freeze/unfreeze feature for scheduling groups
    - deletion has to be explicitly enabled
  - **force recycle bin** for every file to provide a time window to undo unintentional deletion
- Define EC parameters not on directory but space or group level
- Provide availability estimates for each group/space
- Very aggressive file validation (scrub with strong CRC) settings and corruption alarm system

Summary: safe functionality with few add-ons/optimisations

Summary

- Cost difference tape/disk not any more an order of magnitude
- a cold disk storage is very well suitable for online access in terms of access latency if access patterns are mostly sequential - it might be a candidate to merge online and offline storage with small compromises at optimised costs
- EOS can be configured in several ways according to the redundancy requirements today
- EOS can get few extensions to simply setup and monitoring of an erasure encoded storage setup

THANK YOU

QUESTIONS ?

EOS workshop

4-5 February 2019

CERN

Europe/Zurich time zone

There is a live webcast for this event.