Internet Archive
Practical Experience with 8T Archival Drives

John C. Gonzalez

Director of Engineering

jcg@archive.org

The Internet Archive:

Non-Profit Library

Universal Access to All Knowledge

Internet Archive: Quick Overview

- 279 billion web pages
- 11 million books and texts
- 4 million audio recordings (including 160,000 live concerts)
- 3 million videos (including 1 million TV News programs)
- 1 million images
- 100,000 software programs

→ 38 PB of unique storage (mirrored in a ~90 PB cluster)

How?: Some Principles (Reprise)

Transparency	Items = Directories on Disk
Simplicity	Disk = Unit of storage
Durability	Each disk is replicated
Performance at Scale	BOTH disks serve content
Longevity of Access	Evolve formats as needed

For details, see this blog post

Storage Innovations in 2016/2017

- Storage expansion of ~7PB unique (~14PB raw)
- Slight upgrade of “standard 36x disk node” → 32 core 10Gbps
- In compute-centric nodes, shift from 4T O/S drives to small SSD drives (non-stop hot swap use case)
- Deeper experience with Seagate 8T Archival (shingled) Drives
  - Use in General Storage (7164 drives)
  - Use in an experimental back-up (4320 drives filled)
  - Use in HDFS cluster (684 drives)
  - ~38% of total disk population

8T Experience (Since Jan 2016)

- Great $/TB ➔ ~$32 list price (raw, formatted)
- Drive write speeds slow considerably when capacity reaches ~80%
- Took a lot of work to get them stable on our platform
  - without workarounds, drives crash (with latest AR17 firmware), and required power cycling to continue
  - HDFS, with it's heavier write workload, crashed drives more than our archive oriented storage system (both are running same linux on top of vanilla ext4)
  - “Feature” to encourage use of higher-end enterprise class drives??

Future Directions and Concerns

- Introduction of “supplemental SSDs” into storage cluster
  - Incremental compute capability on top of storage role

- 14T and 16T drives are on the horizon…
  - Lot of storage in a single unit (blessing and curse)
  - Even at 10 Gbps, 16T → 3.6 hours to copy

Additional Tech Team Comments

- We love drive managed SMR!
  - Disks looking like an array of blocks means we have minimal software system changes as storage technology evolves
- That having been said, we have no problem with software playing nice with storage.
  - eg: TRIM for ssds and the ideas in ext4-lazyy*
  - Great that these techniques are optional, in contrast to something like host managed SMR file systems

* - https://www.usenix.org/system/files/conference/fast17/fast17-aghayev.pdf

Internet Archive
Practical Experience with 8T Archival Drives

John C. Gonzalez

Director of Engineering

jcg@archive.org

Detail on Firmware issue…

- DATA: 'AR13'->2176 . 'AR15'->874 . 'AR17'->4114
- WBGRP_HADOOP: 'AR13'->258 . 'AR17'->417 . 'RT17'->9
  - We now use RT17 ONLY in HADOOP applications
- Accommodation for AR17: Disable write-back caching