Fixity

Architecting for Integrity

September 2015

LIBRARY OF CONGRESS

Packard Campus for Audio Visual Conservation
http://www.loc.gov/avconservation/packard/

The Problem

"This is an Archive. We can't afford to lose anything!"

- Our customers are custodians to the history of the United States and do not want to consider the loss of data that is likely to happen at some point
- Content is the original submitted data.

Solutions

- At least 2 copies of everything digital
- Test and monitor for the failures / errors
- Refresh the damaged copy from the good copy
- This process must be as automated as possible
- Someday data loss will occur
  - What's that likelihood?
  - What costs are reasonable to reduce that?

Scoping the problem

File Fixity is a digital preservation term referring to the property of a digital file being fixed, or unchanged

http://www.library.yale.edu/iac/DPC/AN_DPC_FixityChecksFinal11.pdf "fixity check"

Fixity checking is the process of verifying that a digital object has not been altered or corrupted

PREMIS 2.0 Preservation Events Collection. Library of Congress Standards & Research Data Values Registry

Fixity is a function of the whole architecture of Archive/Long Term Storage

- Hardware
- Networking
- Software (COTS, Utilities)
- Processes (System admin, logging)
- People
- Budget

Comparing the solutions

The Library invested in a contract to improve our understanding of the relative influence that each of these functions exert on Archive Integrity - the fixity of content submitted by our customers

How much more secure will our customers content be if:

- There is a third, fourth or fifth copy?
- All content is verified once a year versus every 5 years?
- More money is spent on higher quality storage?
- More staff are hired
  - To monitor the systems?
  - To produce standard operating procedures?
  - To test/patch
  - To develop and maintain monitoring utilities?
- Jeff Robinson will be presenting on this

Comparing the solutions

RAID is at risk due to larger disk sizes. How do we protect content on our disk cache and, potentially, on disk archive?

Is erasure encoding a viable alternative?

- RAID _is_ erasure encoding
- What are my choices with erasure encoding?
- Some vendors have a fancy spreadsheet helping me choose how to vary the encoding to accomplish different reliability. What's really going on there?
- Ethan Miller will be presenting on this

Design Principles

- Wide variation in price, performance and reliability
- Performance and reliability are not always correlated with price
- What is your duty cycle? How many GB per day/month/year
- Use the same measures: GigaBytes (1000^3). Remember that most Operating Systems report in GibiBytes (1024^3)
  - GB / GiB: 7.3 % difference
  - TB / TiB: 10 % difference
  - PB / PiB: 12.6 % difference
- Insist on vendors providing failure rates in GB processed
- Choose hardware combinations to limit likely failures based on your duty cycle
  - Disk is rated at UBER of ~ 10^-15 – our duty cycle is 100 TB / month. Every 10 months we are likely to have an UBER