

Designing Storage Architectures for Digital Preservation: The State of the Industry

***Henry Newman
Instrumental Inc
CTO/CEO***

hsn@instrumental.com

Statements about the state



HIGH PERFORMANCE INNOVATION

- Storage is not free and never will be
- Media densities are beginning to have significant growth again after the recession because of new investment
- Software scaling for large archives is a challenge
- The world is not perfect

Storage is not free



HIGH PERFORMANCE INNOVATION

- Contrary to what some industry pundits have been saying storage will not be free or virtually free
 - And it never will be
- Amazon Glacier claims \$.01 per GB per month but read the fine print
 - How are they calculating data integrity?
 - They claim "the annual average data item durability is 99.99999999."
 - Is this the same as data integrity?
 - How is this calculated?
 - How much does it cost to bring it back?
 - How is the retrieval fee really calculated?

- Storage densities are beginning to grow again
 - The recession caused a drop in investment
 - This does not mean that growth will go back the early 90s trends of doubling every 18 months
 - More on this from Robert Fontana
- Is everything going to grow at equal rates?
 - No: Different technologies are hitting different physical limits
- Is the storage reliability going to scale with density?
 - No change over the last 7 years
 - What does that mean for preservation archives?
 - Data loss is not an option

Software scaling challenges



HIGH PERFORMANCE INNOVATION

- Software scaling for:
 - File counts
 - Billions of files in a file system and millions in a directory
 - Data integrity
 - How can integrity be validated from creation to deletion?
 - Migration to new technology
 - Software upgrades
 - Hardware upgrades
 - How do we test these complex systems to meet all of requirements?
 - » This costs time and money!
 - Archive Policies
 - How do you or how can someone delete a file 7 years from now?
 - How does the archive system know it can delete a patient record file x years after the patient dies?
 - Even if an archive system could do this policy, what about migration to other vendors if you want change vendors?
- Archive software is a challenge given policies and lack of standards and there is the light at the end of the tunnel that is not moving at us quickly

The world is not perfect



HIGH PERFORMANCE INNOVATION

- **Stuff breaks**
 - Amazon, Google, Microsoft all have had downtime
 - Has LOC ever been down
 - We all know the answer
- **Security is an on going challenge**
 - What if someone just wanted to change the data not destroy the data
 - In some cases this is more serious than deleting an archive
 - It could be for financial gain, political reasons, or just plain revenge
- **No matter what happens**
 - Who is going to protect us and how from both ourselves and from external forces

- We hope to explore all of these areas and more
- The agenda was developed with the hope that LOC can share their requirements and their concerns
- And everyone can share their wisdom