Instrumental
HIGH PERFORMANCE INNOVATION

Designing Storage Architectures for Digital Preservation

September 27-28, 2010
Library of Congress
Henry Newman

Challenges Libraries Face

Instrumental
HIGH PERFORMANCE INNOVATION

- IT technology changes at different rates
  - Hardware
  - Software
- Preservation community (librarians and archivists) discuss preservation completely differently than IT people and/or vendors
  - Vendors use different “9” counts
  - Librarians use “data loss” or “no data loss”
- Costs
  - The TCO of digital preservation is not well understood
  - Lots of hidden costs and impacts

Hardware Technology Changes

Instrumental
HIGH PERFORMANCE INNOVATION

- CPU speed and cores increasing faster any other component
  - Per core memory bandwidth is dropping as core counts increase
- Memory bandwidth has not been scaling with CPU performance
  - Both needed for data validation and checksums
    - DDR-3 performance has not scaled with CPU performance increases
- Storage performance
  - PCIe bus performance lags CPU and Memory
    - PCIe 1.0 (2003) 250 MB/sec per lane, PCIe 2.0 (2007) 500 MB/sec per lane, PCIe 3.0 (~2010) 1024 MB/sec per lane
      - 8x improvement in 6+ years is just part of the problem
  - Storage connectivity lags CPU and Memory
    - Fibre channel has improved only 4x during the same time
  - Storage devices (**BIG LAGS**)
    - Tapes and disk are far less in terms of performance

Software Technology Changes

Instrumental
HIGH PERFORMANCE INNOVATION

- No standards for checksum management
  - Critical for digital presentation
    - Nothing in POSIX and nothing planned
- No OS standards for increased ECC
  - Since consumer technology is driving the market this is critical concern for preservation
    - ECC has not improved as a function of channel speed
- HSM software has had very limited changes
  - No changes to address digital preservation
    - Requires significant integration of more than just HSM
  - No standards
    - No standards bodies addressing HSM
  - Significant reduction in number of vendors over the last 10 years
    - HSM market is shrinking at least in terms of number of HSM licenses

Example: 100 PB Archive

Instrumental
HIGH PERFORMANCE INNOVATION

- Librarians/archivists discuss preservation completely differently than vendors
  - Discussed in terms of data loss or no data loss
- No one can easily calculate the reliability of data in an archive
  - Best we can do is an estimate
- No vendor nor IT professional will ever provide 100% data reliability for 100 PB of data
- Digital preservation community wants this level of reliability
  - And, of course, at a reasonable cost
    - If they cannot get it they need to know what they can get and the cost

Costs

Instrumental
HIGH PERFORMANCE INNOVATION

- Costs of digital preservation is not well understood
  - Lots of hidden costs and impacts
- 100% data reliability is impossible given the cost for large archives
- What are the actual “9” counts as they relate to cost
  - This is next to impossible to calculate
    - It is not discussed and yet all librarians and archivists want to understand this