Designing Storage Architectures for Digital Preservation
Library of Congress
Meeting Location: Sofitel Hotel Washington, D.C.
Meeting Dates: September 27–28, 2010

Overview

The purpose of the meeting was to bring together technical industry experts, IT professionals, digital collections and strategic planning staff, government specialists with an interest in preservation, and recognized authorities and practioners of digital preservation to identify common areas of interest to inform decision-making in the future. The agenda, a background reading list, and copies of presentations are available at this url:
http://www.digitalpreservation.gov/news/events/other_meetings/storage10/index.html

Opening/Welcome

Martha Anderson, Director of Project Management for NDIIPP at LC opened the meeting, and talked about the changes in these topics over time. When this meeting series commenced the main topic was tape vs. disc. Last year we were talking about resilience. We've learned a lot over the last 3-4 years.

John Warren, IP Solutions, LLC, stated his role is to help with the goals of engaging participants. All the presenters have 5 minutes. Then there will be a free-for-all discussion and Q&A session for each section of the agenda.

Highlights from Steven Johnson on "Where Good Ideas Come From".
Most ideas take a long time to gestate and mature. They come from small hunches. Tim Berners-Lee didn't have a full version of the WWW; it took 10 years for it to incubate. Hunches need to collide to become larger than the sum of their parts. When you look at innovation from this perspective, it's important to remember the big driver for scientific inquiry is the evolution of connectivity. It's true we're more distracted because we have more ways to connect, but chance favors the connected mind.

Overview of Trends in Storage Architecture Solutions

Henry Newman, Instrumental, Inc.

Why are we here? Tech people think of the world in hardware/software. The preservation community thinks in terms of how we keep all this stuff as we move from an analog to digital age. Vendors think in terms of "9" counts. Librarians think in terms of "data loss" and "no data loss". No one really knows what digital preservation costs.

CPUs have become increases fast, but core bandwidth is dropping. PCI Express bus performance lags behind CPU performance. On the software side, there is no standard for checksum management.
Hierarchical storage management software has had little advancement over some time..

There are no end-to-end standards for data integrity and error detection in the complete data path.
There is no standard format for checksums on the vendor side. On the government side, they have been implementing on a format basis. How are these communities going to come together in a cross-domain manner? We've been talking about this for 10 years. When you have a 100PB archive and librarians/archivists are interested in no data loss, how do you estimate the closest you can provide? There is no real way to estimate the reliability over time. So how do you this at a *reasonable cost*, and if not, what does it cost and what sort of performance will you get? You cannot get 100% reliability on 100PB; and so the cost is next to impossible to calculate. Consumers are driving the storage market, not enterprise technology, which has an additional impact on the calculation.

Storage Architectures for Digital Preservation: Case Studies from Heavy Users of Digital Content

Ian Soboroff, National Institute of Standards and Technology—Storing and Processing Big Data

At NIST, we build a research agenda around problems affecting large sets of data (see the blog post from < http://pseudo.posterous.com/storing-and-processing-big-data>).

ClueWeb is a 25TB collection in HTML. For a researcher this is a large dataset, although for a vendor it's small potatoes. You can accomplish a lot with a desktop. Small clusters in the cloud are another possible solution for low-resourced outfits. This is driving storage issues in the research community.

Mark Phillips, University of North Texas—UNT Digital Library

My perspective is from a medium-sized research institution exploring where the digital library world will intersect with traditional library values. UNT had a fairly easy-to-manage library infrastructure with average growth in technology needs, until we began large digital content acquisitions. Now we're faced with how can we continuously scale our storage infrastructure if we don't know how much content we'll ingest over time.

We decided to scale out our SANS, and then we found out how much it cost. We invested in a PetaBox™ and low-end SATA arrays, but they didn't really scale for our needs. It took a long time to build consensus among the various stakeholders we were headed in the right direction, and there were many compromises in the end.

Andy Maltz, Science and Technology Council, Academy of Motion Picture Arts and Sciences—A Digital Motion Picture Archive Framework

In the *Digital Dilemma*[1], one of the predictions for 2015 was significant data loss, which was experienced with the Sidekick event[2].

The Academy has a new pilot system to manage a digital motion picture collection in the Digital Motion Picture Archive Framework Project. This was a learn-as-we-do project, not a long-term preservation solution. We built a 10MB, 3-tiered storage system, with VPN access built on an OS system. Lesson learned were:

- You can't get a degree in archival motion picture management; you have to do it yourself.
- Born archival is important.
- Software can be cheap, but you have to have a crack team.

Ethan L. Miller, University of California, Santa Cruz—Archival Storage Architectures

Pergamum is used in the AMPAS pilot system. Pergamum is a disk-based archival storage with a low-powered CPU on a network. The disks are self-checking; smart so you can ask it questions, and have an idle power load at about ½ a watt. There is parity on each disk, and you can look at erasure codes to prevent data loss.

In parallel, we at UC Santa Cruz we are interested in studying system usage patterns, and are beginning to look at these. We noticed a discrepancy in the way three different storage systems were used. We would like to expand our survey of the usage logs, so please share any data and insights with us.
We are trying to study the questions of how people are accessing stuff, and what they're accessing so we can remove the "probably" issue of building solutions. Figuring this stuff out before building a system would be very helpful.

Raja Rajasekar, University of North Carolina at Chapel Hill—iRODS for CineGrid: Policy-Oriented Data Cloud Management

IRods is used in the AMPAS pilot system.
CineGrid requirements focus on storage and access for distributed data, support for the metadata, and policy-based administration. CineGrid uses the iRODS data system, a cloud of data servers, with a rule engine and integrated metadata catalog. The data lifecycle for the CineGrid project is framed by a federated approach. It boils down to two important factors: 1) definitions of policies, and; 2) definitions of metadata. Retention, disposition, and authenticity are the most important policies being maintained by CineGrid.

Tom Garnett, Biodiversity Heritage Library—The Biodiversity Heritage Library: Preserving a Knowledge Ecology

---

[1] http://www.oscars.org/science-technology/council/projects/digitaldilemma/
[2] http://www.pcworld.com/article/173470/microsoft_redfaced_after_massive_sidekick_data_loss.html

BHL supports current research that uses older resources through digitalization of collections. BHL is at 80TBs today, and is the largest collection of this literature in the world. Preservation requires more than the storage of bits. Access is an aspect of preservation. BHL follows these broad principles:

- Distribute responsibility to larger community
- Increase sense of ownership
- Replicate content in many places
- Distribute to each node, which is self-sustaining.

This is more than a technical process; we at the BHL are focused on the social factors.

Barbara Taranto, New York Public Library —New York Public Library: Digital Storage Requirements

I have heard from people, "if you can't guarantee 100% reliability, then you have no business in this business." Libraries are an evolving ecology. When NYPL went into the repository business, they wanted a high-production workflow attached to the preservation repository. We are in the process of moving TBs of data into the preservation space. Lessons learned for what preservation space needs:

- You need the space ahead of the process.
- Destination space may be larger than staging areas.
- Final space requirements may be larger than the migration space.

The planned-for destination storage is actually smaller than the amount of storage you will eventually need because you need it all over your workflow. If not, you end up with a lot of garbage you can't use taking up valuable space.

Martin Kalfatovic, Smithsonian Institution Libraries—Creating a Digital Smithsonian

SI recently released a roadmap for how they're going to digitize all of its resources for the widest use. SI plans on digitizing a wide variety of data, including geo-sensing data. Born-digital art collections are streaming into SI more increasingly. We are faced with these questions:

- What does it mean to digitize an object in the collection, e.g., a locomotive?
- What are our storage requirements?
- What do we collect, and how will it all come together?
- How do we aggregate metadata from a variety of collections and systems?

Tab Butler, MLB Network—IT and Digital Workflows: Our National Pastime, All the Time

MLB.com has been online since 2008. We had 6 months to hit 55 million households on day one. Content drives our infrastructure design, which is used for many different purposes. All the information is managed by an asset management system, built on top of a tape storage system. For every hour of baseball, there's 5 hours of content. We put together a workflow using SAMMA robots. Eighteen loggers work daily with the digital asset management system. We have used about 9,000 tapes in 2010 alone.

Storage Architectures for Digital Preservation: Case Studies from Community Service Providers

Cory Snavely, University of Michigan—HathiTrust: A Shared Digital Repository

My group runs storage and data ingest for HathiTrust. Most of the content is Google-scanned materials, and we have some locally-digitized content from partners. Our process got faster with improvements in infrastructure, which takes advantage of the architecture of system itself. Archiving is fundamentally about a system administrator's worries. Storage system vendors can provide a better solution in some instances. Data integrity solutions are important.

Bradley McLean, DuraSpace—DuraCloud Pilot Program: Experiences

Duracloud utilizes the cloud infrastructure for preservation support, yet there are little to no SLAs available for the underlying providers. We're trying to build services and capabilities on top of cloud storage. The evaluation of the pilot will be available soon. We currently have about 10TB from three partners, and learned a lot about moving data back and forth, chunking, and stitching things back together. We definitely had the local copies problem. With 42 processes across 6 servers, bandwidth is a problem. Large files are challenging and naming matters. I would love to see a standard around checksums.

Matt Schultz, Educopia Institute—Disaster and Contingency Planning: Storage Dimensions

Educopia is a distributed network of multi-TB servers with a private LOCKSS network infrastructure. The cloud is working well for the MetaArchive Cooperative, and we are currently testing interoperability with the Chronopolis project. We initiated a dry run of a disaster recovery process in which we took a West Coast cloud image of the East Coast server network, which happened in less than 6 hours. Amazon S3 is meeting the needs for geographic distribution. Cloud storage hasn't improved in costs for full system administration needs. Contingency planning was more involved for us, and required some adjustments.

David Minor, San Diego Supercomputer Center—SDSC Storage: New Developments

SDSC is going through a re-implementation of their data center, looking at opportunities to start from scratch based on user needs. We've seen a movement from pure computation to data-driven research, and are trying to tie our hardware stack to three tiers of users. Our environment supports high-performance computing storage, traditional fileserver storage, and archival services. We have interrelated services and layering: provided natively, which layer well with software and funding, and are shared with different architectures. Data Oasis SLAs and agreements will see a lot of focus, because the guarantees are important. We are moving away from tape-based system to disc-based system, and are eagerly awaiting the NSF data-management policies.

Data Integrity: Ensuring One Good Copy

Micah Beck, University of Tennessee—An Argument for Lossy Preservation

Here's a characterization of what I see as the current art in preservation: bit preservation is possible without so much loss you have to worry. The approaches are decentralized, distributed, set up as wide-area systems, or "tapes in a cave" like Portico. Regarding scalability; how much data can we store in these ways, and for how long? What if lossless preservation won't scale? There are people who want to keep everything: all the data coming off security networks and earth-sensing systems, for example. If we can't do lossless preservation in a scalable way, then we may have to settle for lossy preservation. How sure are you we can do lossless preservation? My research approach is using high-level structure for looking at corrupt files. We already do this with film.

Mike Vandamme, Fujifilm Storage IQ—Data Integrity: Tape Archive Verification

More than 60% tapes pass tests, but there is a lack of visibility in backup applications. Fuji's ReadVerify Appliance reads data before problems creep into backup apps. It hooks up into a SAN switch, and gives you a view into the backup system. Every tape and disk can be traced, and performance tracked. You can maximize and balance load through active management. The ArchiveVerify feature allows you to automate tape reading without accessing the backup in order to verify the old tape archive at near-streaming rates. It uses SCSI VERIFY command to access the drive. You can get an overview of your environment in a report.

Hal Woods, HP [no slide presentation]

What is forever? Should this world come to an end and an alien civilization landed here, would they be able to read our entire genealogical record? Technology is not your friend, but there is a predictability pattern. We can't predict growth rates, etc. I get to speak to many markets and found some best practices across these markets:

- Data loss through reference is a problem, e.g., link rot. Do you take a snapshot of the entire web when you have links?
- Every time something is installed, something is retired.
- Develop an architecture that grows in place, and retires in place. The infrastructure needs to evolve over time.

Mootaz Elnozahy, IBM—Trends in Reliable Storage: A Guide to the Intelligent Buyer

- Interaction with power—you will need power management because all storage systems will have some aspect of this, and you're already beginning to see signs of this in large-scale storage. The way disks are designed today leaves them vulnerable to power shutdowns.

- System & problem size—storage density won't keep up with growth, precipitating a need for more components, which will lead to more failures.
- Hardware trends in soft errors—we have seen orders of magnitude increases in soft error rates, and we'll have to ask tough questions about architecture. You need to understand how power management is done.

Dave B. Anderson, Seagate Technology—Hard Drive Directions

Growth in hard drive capacities is plotted by a series of S curves. The key to your market is notebook 2.5" drives. Even desktops are coming out with 2.5" instead of 3.5" drives. High-capacity storage is still dominated by 3.5" drives, but 2.5" is making in-roads. SSD is still a very small market. You'll see 3TB drives come out in 2011, which breaks the 4-byte address. In the foreseeable future, most of your storage will be on magnetic storage devices. When things tip towards 2.5" drives, you'll get much more storage for the dollar.

Mike Smorul, University of Maryland—Audit Control Environment (ACE)

How do you manage a fixity digest at a software level? ACE software was developed at the University of Maryland. The idea is that the software issues a small token that sits along side of data and is cryptographically linked to your data. The ACE Audit Manager resides as a local archive, and is freely available online. Most digests will claim a file is intact, but have no temporal context for when the file was intact. That's where the token's usefulness comes into play. Witness values can be pulled from a trusted source to provide information. This is working with Chronopolis currently.

Data Compression and De-duplication Trends

Joe Zimm, EMC2—Data Domain De-duplication Overview

De-duplication has mostly been done on backup. This gets various levels of compression; whitespace reduction, file level, fixed blocks, and variable segments and local compression. Whatever reduction we get locally, is replicated across a network. We generally get a 50–60% compression rate on the second full backup. Traditionally this has been used in backup environments, but now is being used on archived datasets. This reduces the footprint and power requirements.

Mike Davis, Ocarina/Dell—Trading CPU Cycles for Gigabytes: Data Reduction Approaches for Archival Storage

Everyone has collections that come in sporadically and are working under unstable budgets, so how do you reduce storage costs? You can move your data around more easily when compressed. Our product does both de-dupe and compression. Can you add compression to already compressed data? Yes, you can de-layer objects. It's not just about algorithms. Compression is core intensive; de-dupe is RAM intensive. 80% of access to storage is metadata driven, so we want to keep this information intact. Maybe we should consider lossy compression early in the lifecycle.

Steve Vranyes, Symantec [no slide presentation]

De-dupe is all the rage over the last year. In the next few years, you're likely to see more backup environments becoming smarter. At Symantec, we're seeing better value in compression rather than in de-duplication. De-dupe is no longer a specialty thing; you'll see it more as a built-in commodity underlying more vertical products. What's going to matter in the future is scale. What matters most in de-dupe is memory. Compression will allow must better results for the future.

Raghavendra "JP" Rao, Cisco—Data Compression and De-duplication

We're memory-constrained in our environment. The main goal is to improve WAN optimization. Wide area application service, WAAS, is being used for replication. We're creating a data-agnostic environment.

Processing and Analytics of Large Datasets

Jimmy Lin, University of Maryland [no slide presentation]

I've been working with Twitter, and we're observing 150 million users. Twitter believes in an open-source stack, and they want to run analytics on top of that data. At Twitter it's hard to find the right questions, but easy to get answers. Where are the feeds coming from? What are the characteristics of a tweet?

Mike Smorul, University of Maryland—Web Archive Processing

We're processing web archives at the University of Maryland with little budget, using management tools and content from the Library of Congress. Webarc sits on top of a MySQL database. It has a REST-API, and a Javascript client. We developed a Simple, Web-Accessible Preservation (SWAP) application, which performs fairly well. We're working on a time machine for the Web, tackling how to present data in a way a user can understand it. We need to scale out to 32 nodes.

John Johnston, Pacific Northwest National Laboratory—High Performance Computing for Data Intensive Science

We focus on data-intensive computing, which includes computational science and multi-media analytics. There has been a massive deluge of data in the last 15 years, as computation has become an integral part of the scientific method. The risk to traditional high-performance computing and data-intensive HPC is they have the same starting point; that you will need data and it won't be there. You have to enable researchers from their desktop computers.

Leslie Johnston, Library of Congress [no slide presentation]

The IBM Emerging Technologies group has been working with the British Library. IBM then came to LC with the Big Sheets solution, to put structure on 180TB of unstructured data. The tool allows you to ingest the data, process it, and run some analytics on it. The Library of Congress Web archives is a selective one, mostly topical collections, including five sets of recent elections that we want to treat these as a single, longitudinal data collection. The Library provided 8TB of data to run some term-frequency analysis, name-entity analysis, and basic visualizations. We're working with ARC and WARC formats, and the IBM tool allows for these packaging formats. Bag-It support is planned. There is an example from the British Library available from http://www.webarchive.org.uk/analytics/analytics.htm.

Dave Fellinger, DataDirect Networks—Reducing Service Latency in Large Storage Systems

We have supported much unstructured data like video and photos in ShutterFly. Bottleneck analysis on the hardware and software chains tells a story about latency. With many SCSI layers and bus transitions, and bandwidth issues, many consecutive steps can cause issues. System simplification through virtualization can eliminate multiple layers. Think about file-reads/sec rather than IOPs. Don't be afraid to count your I/Os, because you pay for them.

Kevin Kalmbach, Sun/Oracle—Storage Directions

Processors are getting faster and storage isn't. The cheap and ubiquitous processors are falling into storage systems. Now there's a lot of capability to do a lot of heavy lifting in storage systems. So where do you manage what? Oracle came out with ZFS, a 128 bit, open-source filesystem. It does RAID holds at the file level, and end-to-end checksumming, so you know about bit rot when it happens. Tape drives are getting larger, and the context matters when you think about really large and disperse datasets.

Format and Technical Migration

Subodh Kulkami, Imation—Format and Technology Migration

We deal with many types of data storage options. There will be a slow migration in capacity with magnetic and optical solutions, and more significant increases in marketshare for flash and hard disk products.

David Rosenthal, LOCKSS—How Green is Digital Preservation?

This content is described in more detail at:
http://blog.dshr.org/2010/09/how-green-is-digital-preservation.html
Vijay Gill from Google showed that most of the cost to run a data center is spent on power and cooling. Even though the carbon footprint is large, additional data isn't affecting the picture much at the moment. Data on disk is moving away from the CPU to

small clusters. These will survive benign neglect very well because they plug into a node with power over Ethernet.

URL's Referenced During the Meeting

Steven Johnson – Where Good Ideas Come From
http://www.youtube.com/watch?v=0af00UcTO-c&feature=related

The following websites were referenced during a discussion of repositories of technologies and documentation that might be useful for digital preservation.

Computer Failure Data Repository
http://cfdr.usenix.org/.

National Media Lab
http://nlm.org

Preserving Virtual Worlds project
http://hdl.handle.net/2142/17097

Photo Metadata
http://www.photometadata.org/

Digital Photography Best Practices
www.byBestflow.org

Charles Babbage Institute
http://www.cbi.umn.edu/