Why Traditional Storage Systems Don't Help Us Save Stuff Forever

Mary Baker
HP Labs, USA
mgbaker@hp.com

Kimberly Keeton
HP Labs, USA
kimberly.keeton@hp.com

Sean Martin
British Library, UK
sean.martin@bl.uk

Abstract

We are in the midst of an unprecedented transformation from physical to virtual assets. Online contracts, digital photographs, digitized movies, music, technical journals, corporate records, web sites, and government documents are just a few examples of valuable digital assets that organizations would often like to preserve for long periods of time – not just for years, but for decades or even forever. Unfortunately, long-term preservation remains a huge challenge due to the unusual nature of the threats from which it suffers compared to traditional (shorter-term) storage applications.

Our goal in this paper is to describe how these environments differ and to acquaint the dependability community with some of the challenges in building archival storage systems. We give some guidelines for an alternative storage architecture, much of which is being implemented at the British Library, and we conclude with some suggestions for initial research topics to be tackled in this area.

1. Introduction

Most research and development efforts in the large-scale storage area concentrate on the goals of traditional enterprise storage systems. This is entirely reasonable given that most sales of large-scale storage systems are for enterprise environments. However, we believe another important storage area with different characteristics and requirements is emerging as a result of the growing digitization of previously analog assets. Corporate auditing regulations (such as Sarbanes-Oxley [1]), and new methods of creating and capturing content (such as digital photography and online publishing) are reasons why we increasingly have digital assets worth saving for long periods of time. Unfortunately, our large-scale storage architectures are generally not helpful for long-term preservation of online content.

In this paper, we contrast the requirements for large-scale enterprise storage systems with the requirements for archival storage systems. We list the threats that apply to archival storage systems and why they make building such systems a challenge. We describe some guidelines for solutions to these problems and provide a case study of a digital archival system at the British Library that is incorporating some of these solutions. However, this is only the tip of the iceberg for research into long-term reliability of data and large-scale archival storage systems. Many of our ideas and potential solutions remain untested, and we look forward to the results of current and future evaluation efforts.

2. How Enterprise Storage and Digital Preservation Applications Differ

In this section we consider qualitatively how the goals and architectural requirements of traditional enterprise storage systems differ from archival storage systems. We contrast a transaction processing enterprise system with a large-scale library archival system.

The goal of traditional enterprise storage systems is to meet customer-specific or application-specific needs for performance, availability and reliability. While these characteristics do differ across customers and applications, the architectural requirements are often demanding. For instance, a transaction processing system may require high read and write performance (in terms of both latency and bandwidth), almost no downtime, and reliability that limits the number of recent transactions that can be lost to only a few minutes' worth.

In contrast, the goal of long-term archival storage applications is to preserve and provide read access to stored content for however long it is valuable – perhaps forever. Thus it is more important that the systems be reliable and available in the long term

rather than on a minute-by-minute basis. Another key requirement for archival systems is low cost, as there is often little budget for preserving old materials into the future. Huge capacity is also needed for archives that want to bring online hundreds of terabytes or even petabytes of data per year. However, it is not just current capacity that is important but the ability to scale tremendously over long periods of time.

While performance matters for archival systems, we believe the workload and access patterns differ from enterprise applications. Based on the small amount of workload information so far available from online repositories (such as the British Library), and extrapolating from access patterns of large traditional repositories, we believe that read accesses comprise the bulk of the workload. In addition, accesses are generally spread across a large body of content with little locality, so the access probability for any particular item remains very low, providing natural load balancing across the archive. In contrast, many transaction processing workloads have high access locality for parts of the data set, the metadata or the indices, which allows (and requires) designers to exploit caching to improve performance.

Write accesses in archival systems are generally confined to ingestion of new material and migration of data from old to new formats or from old to new media. The invariance of stored objects means that there are almost no updates in place. There may be some re-purposing of material or re-signing of content to support its authenticity, but generally this creates new versions of the material. In contrast, transaction processing workloads may have significant update traffic to data, indices, and logs.

Other important differences in requirements come to light when we consider the lifetimes of the storage systems. A customer may expect to purchase the bulk of an enterprise storage system from a particular vendor with the expectation that the vendor will provide service for the lifetime of the system, which might be estimated at 5-10 years. These systems tend to be decommissioned after a new system has been brought online. Although the storage system is not expected to last forever, considerations of how data will be migrated from the system at the end of its lifetime are rarely part of the purchase decision.

The long lifetime and scalability requirements of an archival storage system mean that the entire system cannot be purchased at once. Instead, capacity must be scaled up over time with additional purchases, while hardware at end-of-life is decommissioned over time. It is not affordable to decommission the entire system and bring online an entirely new system. This "rolling procurement" and "rolling replacement" mean that the systems are necessarily heterogeneous, including components with different technologies (media, product generations and interfaces) from different vendors at any point in time. The data the system needs to store must last longer than the lifetime of any of the storage components, any of the storage technologies, or potentially even the lifetime of the companies selling and supporting the storage products. It is thus essential to know that the data can be moved forward through new procurements over time and easily extracted from old portions of the system.

In summary, the key drivers for enterprise systems are performance, availability and short-term reliability. The key drivers for archival storage systems are data longevity, low cost, and scalability over time, technologies and vendors.

3. Threats to Digital Preservation

It is not just the goals and characteristics that differ between enterprise and long-term preservation systems. Due to the long expected lifetime of digital preservation systems and their need for complete reliability, the nature of storage failure threats differs as well. In this section we categorize the threats to long-term preservation of digital assets and explain where these threats also apply to enterprise storage systems and where they do not. Some of the threats only become problematic when assets need to survive for very long periods of time. We label each of the threats appropriately as "HW/SW," "environmental," "people," and "institutional." Some can be caused by several of these sources.

Massive storage failure (HW/SW, environmental, people): Even expensive storage systems can fail, losing large amounts of data. There are many possible reasons for such failures, including compounded or cascading hardware and software failures, natural disasters, human error and acts of war. If the data cannot be restored or regenerated, then it is not preserved. Massive storage failure is also a problem for short-term storage applications, but the likelihood of its occurring is greater over the longer desired lifetime of archival assets.

Mistaken erasure (people): One of the common ways in which data is lost is through users and operators accidentally deleting or overwriting content they still need, or accidentally or purposefully deleting data for which they later discover a need. Like massive storage failure, mistaken erasure also threatens short-term assets, but its likelihood increases with the lifetime of the assets.

Bit rot (HW/SW): No affordable digital storage medium is completely reliable over long periods of time, since the medium may degrade, resulting in "bit rot." For instance, recent studies [6, 8] indicate that CD media – popular with home users and small businesses – are often only reliable for 2 to 5 years, not decades as advertised. Other media such as disks and magnetic tape also can suffer from bit rot. Bit rot can also refer to other *undetected* storage failures that change retrieved content, such as errors in the network interface, software buffer overruns in the operating system, error correction failures in memory, and so forth.

Outdated media (HW/SW): Over time storage media, such as nine-track tapes and punch cards, become outdated. Bits stored on these media become useless when appropriate media readers are unavailable. This problem is largely specific to removable media, where the medium on which content is stored can be separated from its reader [7]. A recent inventory of a local electronics store revealed that it is now difficult to buy an off-the-shelf PC with a built-in floppy drive. Only a few years ago, floppy disks were the lowest common denominator for storage. This problem is less common for short-term storage applications, since they often do not outlive the utility of the medium on which their content is stored.

Outdated formats, applications and systems (HW/SW): In a similar way, application formats become obsolete. Bits stored in these formats (such as WordPerfect for the Mac) become unusable when appropriate applications are no longer available to interpret the content. As old formats become endangered, long-term preservation requires refreshing of data into new formats or perhaps "universal" formats (formats that can represent everything and whose specifications everyone has). A potentially costly alternative is to emulate the platform on which the old application ran, continuing to manipulate old application data in the emulated environment. Unfortunately, it can be hard to capture the entire original environment; enterprise applications, for example, may exhibit dependence on external license servers. This problem is rare for short-term storage applications, since they usually do not outlive the applications that interpret their data.

Loss of context (HW/SW, people): Metadata, or more generally "context," includes information about layout, location, and inter-relationships among stored objects, as well as the subject and provenance of content, the access controls, and the processes, algorithms and software needed to manipulate that content. Preserving contextual metadata is as important as preserving the actual data, and it can be very hard to recognize all required context in time to collect it. A particularly challenging example is encrypted data, since preservation of the decryption keys is essential alongside preservation of the encrypted data. Unfortunately, over long periods of time, secrets (like decryption keys) tend to get lost, leak or get broken [3]. Access controls often become meaningless or incorrect as the listed users change roles or the listed roles cease to exist. This problem is less of a threat to short-term storage applications where the assets do not live long enough for secrets to be lost, for the context to be lost, or for the context to become uninterpretable.

Attack (people): Online repositories are prone to destruction, censorship, and modification of stored data; access disruption through denial of service attacks; and theft of data and storage devices. The attacks may be short-term or long-term, legal or illegal, and motivated by ideological, political, financial, or legal factors, as well as simply a challenge for unbalanced minds. Evidence from the experience of traditional libraries [10] suggests that well-organized and well-funded groups will seek to remove, destroy or alter as many copies as they can access of works with which they have ideological differences. While attack is a threat to short-term storage as well, researchers usually focus on short-term, intense attacks rather than long-term, slowly subversive attacks.

Budget (institutional): Many organizations with materials to preserve do not have large budgets to apply to the problem and hope to declare success after just managing to get a collection put online. Unfortunately, this provides no plan for maintaining a collection's accessibility or quality in the future. Motivating an investment in preservation can be difficult [4] without better tools to predict long-term costs, especially if the target audience for the preserved information does not exist at the time decisions are made. Although budget is an issue in the purchase of any storage system, it is usually easier to plan how the costs will be amortized over the lifetime of a system that does not need to grow indefinitely, migrating through new technologies and administrative techniques.

Organizational failure (institutional): Organizational failure rarely enters into the design of conventional storage systems, but it can be a big threat in digital preservation. Organizational failure is the dependence on a single sponsoring organization, a single administrative domain, a single vendor, or a single service provider. For instance, it is quite likely that preservation of an online archive will fail if the organization sponsoring it goes out of business or changes its priorities. Assets that must live for long

periods of time must be protected from the failure of any one organization, administrative domain, vendor or technology. "Exit strategies" for vulnerable data need to be better understood and promoted.

4. Architectural Solutions

In this section we outline some initial mechanisms for addressing the requirements of archival storage systems. The differences in architecture center around the need (as compared to enterprise storage systems) for lower cost over time, lower vulnerability to organizational failure, lower vulnerability to long-term attack, and higher long-term reliability. Fortunately, archival systems also have weaker requirements on read access performance and short-term availability. The solutions focus on replication across autonomous sites, lower per-site engineering requirements, and the ability to scale over time and different technologies.

4.1 Replication across autonomous sites

For long-term preservation of data that must not be lost, geographic replication of the data across administrative domains is essential for several reasons. First, replication allows data to be preserved across a site failure. Second, it makes it possible to audit the content at one site against content at another site, through a voting process on content digests or another comparison mechanism [9]. This auditing allows us to detect corruption of content due to bit rot, attack and human error in environments where trusted local auditing techniques are not available or are not sufficiently reliable. Third, the sites can be autonomous in administration and diverse in technology, providing increased resistance to organizational failure, human error and attack.

Minute-by-minute consistency between sites is unnecessary for long-term preservation, so the sites can be loosely coupled. This means we can use application-level consistency techniques, rather than expensive solutions such as the inter-array synchronous mirroring required in some enterprise environments. Content should be mirrored across sites, but the structure of the data need not be. The technologies deployed (both hardware and software), the arrangement of data, and the administrative techniques used can, and should, vary considerably across sites. This autonomy of the individual sites reduces their vulnerability to many types of failures.

In contrast, many enterprise systems may consider building a highly reliable, high performance single site. This consolidation of defenses and resources is very attractive, but it does not provide the threat resistance that is essential for long-term digital preservation.

4.2 Low per-site engineering costs

The replication described above is also useful in reducing the engineering requirements of each individual site. For instance, site replication may be sufficient to recover from local failures, allowing individual sites to forgo local data protection techniques, such as snapshots and backup. In the event of loss, data can be restored from a remote replica. The potential reduction in capital and operational costs for managing backup is extremely important for large long-term archives, as operational costs over time must be very low.

Further, the weaker requirements for read access performance and the lack of locality of access allow for a less expensive system. The relaxed access latency requirements mean that the storage system may be built out of commodity components, rather than high-end state-of-the-art storage devices. The lack of locality in access patterns means that there are naturally no hotspots in the storage system, and so load is effectively spread over the storage devices without the need for direct management. Popularity-induced hotspots are short-lived, relatively small in size, and can be handled by a front-end web server. (Such a hotspot occurred when the British Library announced their digitization of 20K pages' worth of original Shakespeare manuscripts, for instance.)

4.3 Design for long-term scalability

As new materials and new collections are ingested, an archive must scale over time. To prepare for a new collection stream, an archive may purchase new storage capacity. Over time, the storage technology that offers the lowest cost per capacity will change, so new procurements may differ in technology from other parts of the system. Designing for long-term scalability thus means accommodating heterogeneity in the system and avoiding vendor or technology lock-in.

To accommodate heterogeneity, new additions must integrate well with the current system and allow it to grow correspondingly in the future. For example, we must allow for large variations in performance and delay across different portions of the system. CPU performance of different components might vary by an order of magnitude, while the memory performance of

the fastest components might be several times that of the slowest components.

Avoiding vendor or technology lock-in is critical for archival systems, since the archives must last longer than any storage technology or storage company is likely to last. This issue affects system design in the choice of interfaces to storage components. For instance, we might achieve higher performance through a low-level technology-specific interface, but we are more likely to interoperate easily with past and future components if storage is accessed through a few high-level, standardized interfaces (such as a file system interface). There are at least two reasons for this belief. First, a high-level interface allows technological innovation in the storage devices without requiring changes to the interface itself. Second, higher-level interfaces tend to evolve more slowly, especially when they are not vendor-specific.

Fortunately, the workload of such an archive also improves our chances of growing it gracefully over time. The random nature of accesses makes it possible to add new materials and new storage capacity to the system without having to rebalance the existing content.

5. Case study – The British Library

In this section we briefly describe the system architecture deployed by the British Library Digital Object Management Programme (DOM) – a large archival storage system that has made many of the architectural decisions suggested in this paper.

Created in 1972 from older archival institutions, the British Library is a heavily-used world-class research library. Since 2003, it has been legally required to collect and preserve non-print (digital) materials. The mission of the Digital Object Programme is to enable the UK to preserve and use its digital intellectual property forever.

The current target size of the DOM archive is over 800 TB. Because the collection will be amassed over time, flexible, scalable procurement is essential. Additional design goals include inherent scalability in terms of capacity, the number of objects, and the ability to deliver objects. Their goal is to preserve intellectual property forever, so the likelihood of object loss must be infinitesimally small.

Investigations into commercially available products revealed no ready-made solutions sufficient to support their needs at an affordable price, as evidenced by briefings from over 30 leading storage vendors. As a result, the archive's architecture is being designed internally, with several design principles in mind. Disaster tolerance is provided by a multi-site solution. Furthermore, because the system will evolve over time, the logical architecture must support successive generations of physical architectures.

Their design is thus based on replicating content across multiple autonomous peer sites where all sites are active during normal operation. Short interruptions and some degradation in service can be tolerated, but extended loss of complete service cannot. The multi-site solution provides this level of availability without further engineering costs. The system can tolerate the failure of an individual site for a significant period of time, for example, while a replacement site is procured after a disaster. This is because a gateway in front of the site will direct requests to an available remote site, perhaps with some degradation in service if, for example, the requests must travel over a wide-area network.

Cost is a key driver, so they depend on their multi-site solution to eliminate the need for local backups and other expensive techniques for increasing local resilience. They apply local techniques such as RAID only as affordable commodity products that require very little ongoing attention from operational staff and that reduce the probability of needing to pull large amounts of data over the network in the event of a site failure.

6. Caveats

In this paper we have oversimplified the area of enterprise storage solutions. Clearly they vary according to application and customer requirements. The values we ascribe to these systems may be an extreme point, although useful for our comparison purposes. In fact, we believe that many of these enterprise applications will increasingly share some degree of requirements with long-term preservation systems. In the enterprise space, trends in utility computing suggest it is becoming too expensive to decommission large systems in their entirety. In addition, vendor lock-in and other archival threats may become more apparent in the enterprise space.

It is also the case that not all large archival institutions believe in our position. Some are building centralized systems with high availability requirements [2, 5]. They must engineer, even over-engineer, these archives carefully to provide the desired level of reliability and availability within the confines of a single administrative system. Very few digital preservation projects have the funds to accommodate these solutions.

7. Conclusions

In this paper we have argued that there is a new and tangibly different storage area for which traditional solutions are not applicable. The threats to digital preservation make designing archival storage solutions challenging, but we believe the time is right for researchers to apply themselves to this problem. Here we include an initial list of potentially fruitful research avenues that have not yet been sufficiently pursued by the dependability community.

- Understanding how an increased requirement for data longevity (addressing the threats listed in Section 3) affects cost and other dependability axes such as availability, security, reliability, and performance.
- Investigating reliable frameworks that lend themselves to rolling procurement, as described in Section 2.
- Investigating the best means of auditing content across sites (for detecting corruption and other failures) for varying numbers and scales of replicas.
- Automating approaches to data format conversion and evolution of access controls and other metadata.
- Developing fault injection techniques for studying threats to long-lived data. For instance, how do we measure the robustness of systems against low-grade but long-term attacks [9]?
- Characterizing archive access patterns. Much of what we believe to be true about online archives is based on access patterns for paper archives. We want to understand whether these characterizations remain accurate for online content, as well as how behavior may vary from one repository to another.

Undoubtedly, many new research topics will become apparent over time.

8. Acknowledgments

We gratefully acknowledge the help, advice and ideas of the LOCKSS project members, especially David Rosenthal and Vicky Reich. We also thank our colleagues at HP and the British Library for their support and comments. Of course, any errors are solely the responsibility of the authors.

9. References

[1] http://www.techlistings.net/xlist/tech/bizsoft/compliance/sox, 2004.

[2] UC Libraries Preservation Repository: System Design. Tech. Rep. Version 1.8.8, California Digital Library, University of California, Feb. 2004.

[3] Diffie, W. Perspective: Decrypting the Secret to Strong Security. http://news.com.com/2010-1071-980462.html, Jan. 2003.

[4] Gray, J., Szalay, A., Thakar, A., Stoughton, C., and VandenBerg, J. Online Scientific Data Curation, Publication, and Archiving. Tech. Rep. MSR-TR-2002-74, Microsoft Research, 2002.

[5] Harvard University Office for Information Systems. Digital Repository Service. http://hul.harvard.edu/ois/systems/drs.

[6] Horlings, J. Cd-r's binnen twee jaar onlees-baar. http://www.pc-active.nl/toonArtikel.asp?artikelID=508, 2003. http://www.cdfreaks.com/news/7751.

[7] Keeton, K., and Anderson, E. A Backup Appliance Composed of High-capacity Disk Drives. *Proceedings of the Workshop on Hot Topics in Operating Systems (HotOS VIII)*, May 2001.

[8] Malda, R. The Myth of the 100 Year CD-Rom. http://slashdot.org/article.pl?sid=04/04/22/1658251&mode=flat&tid=137&ti, 2004.

[9] Maniatis, P., Roussopoulos, M., Giuli, T., Rosenthal, D. S. H., and Baker, M. The LOCKSS Peer-to-Peer Digital Preservation System. *ACM Transactions on Computer Systems 23, 1* (2005).

[10] Reich, V. Stanford libraries. Personal Communication on the Comparison of Risks between Physical and Digital Assets, May 2004.