David S. H. Rosenthal

What Could Possibly Go Wrong?

DOI 10.1515/bfp-2015-0022

Abstract: The LOCKSS[1] Program at the Stanford University Libraries has been building tools for libraries to use, to collect and to preserve material published on the Web for nearly one-sixth of a century. While there are no one-size-fits-all solutions to the problems of digital preservation, many of the lessons learned during that time are applicable to other types of content.

Keywords: LOCKSS; threat models; long-term digital storage

Was kann schon schiefgehen?

Zusammenfassung: Das LOCKSS Programm an den Stanford University Libraries entwickelt seit fast 61 Jahren Werkzeuge für Bibliotheken, damit diese elektronische Publikationen sammeln und erhalten können. Obwohl es keine Einheitslösung für die Probleme der digitalen Langzeitarchivierung gibt, sind doch gewonnene Erfahrungen auf andere Inhalte übertragbar.

Schlüsselwörter: LOCKSS; Risikomodelle; Digitale Langzeitspeicherung

1 The LOCKSS System

In the paper world, librarians have two responsibilities, to provide current scholars with the materials they need, and to preserve their accessibility for future scholars. They do this through a massively replicated, loosely coupled, fault-tolerant, tamper-evident system of mutually untrusting but cooperating peers that evolved over centuries. Libraries purchase copies of journals, monographs and books. The more popular the work, the more replicas are in the system. The storage of each replica is not reliable; libraries put them in the stacks and let people take them away. Most times they come back. Losses can be repaired via inter-library loan and copy. There is a market for replicas; as the number of replicas of a work decreases, the value of a replica in this market increases, encouraging librarians with a replica to take more care of it, by moving it to more secure storage. The system resists attempts at censorship or re-writing of history precisely because it is a loosely coupled peer-to-peer (P2P) system; although it is easy to find *a* replica, it is hard to find *all* the replicas, or even to know exactly how many there are. Although it is easy to destroy a replica, it is hard to modify one undetectably.

Almost one-fifth of a century ago Stanford's HighWire Press pioneered the transition of academic journals from paper to the Web when they put the *Journal of Biological Chemistry* on-line. This destroyed two of the pillars of the paper system, ownership of copies, and massive replication. In the excitement of seeing how much more useful content on the Web was to scholars, librarians did not think through the fundamental implications of the transition. The system that arose meant that they no longer *purchased* a copy of the journal, they *rented* access to the publisher's copy. Renting satisfied their responsibility to current scholars, but it couldn't satisfy their responsibility to future scholars.

Librarians' concerns reached the Mellon Foundation, which funded[2] exploratory work at Stanford[3] and five other research libraries. Stanford contributed the LOCKSS system[4], which already existed in prototype form thanks to a small grant from Michael Lesk at the NSF (National Science Foundation). It was based on two insights:

- The paper library system was a system in the physical world that had a very attractive set of fault-tolerance properties.
- An analog of the paper system in the Web world could be built that retained those properties.

[1] Lots Of Copies Keep Stuff Safe. LOCKSS is a trademark of Stanford University.

David S. H. Rosenthal: dshr@stanford.edu

[2] Rosenthal, D. S. H.: A Brief History of E-Journal Preservation. In: DSHR's Blog from August 17, 2011, http://blog.dshr.org/2011/08/brief-history-of-e-journal-preservation.html (last downloaded 2015-01-07).

[3] Rosenthal, D. S. H.: It was fifteen years ago today. In: DSHR's Blog from October 4, 2013, http://blog.dshr.org/2013/10/it-was-fifteen-years-ago-today.html (last downloaded 2015-01-07).

[4] Rosenthal, D. S. H.; Reich, V.: Permanent Web Publishing. In: Proceedings of the FREENIX Track: 2000 USENIX Annual Technical Conference, San Diego, CA, June 18–23, 2000, pp. 129–140.

By analogy with the stacks, libraries would run what can be thought of as a persistent Web cache. The cache would be pre-loaded by a Web crawler which would crawl the content to which the library subscribed, and would never be flushed in normal operation. The contents of each cache would be monitored by a P2P anti-entropy protocol. Any damage it detected would be repaired by the Web analog of inter-library copy. Because the system was an exact analog of the existing paper system, the copyright legalities were very simple. The Mellon Foundation, Sun Microsystems and the NSF funded the work to discard the prototype and build a production-ready system.

The prototype's anti-entropy protocol had gaping security holes. It was replaced by a real P2P anti-entropy protocol, which won Best Paper at SOSP[5] a tenth of a century ago. I have long thought that the fundamental challenge facing system architects is to build systems that fail gradually, progressively, and slowly enough for remedial action to be effective, while emitting alarming noises to attract attention to impending collapse. The interest in our paper is that it shows such a system, albeit in a very restricted area of application. It is a true P2P system with no central control (which would provide a focus for attack), using three major defensive techniques:

- Effort-balancing, to ensure that the computational cost of requesting a service from a peer exceeds the computational cost of satisfying the request. If this condition isn't true in a P2P network, the bad guy can wear the good guys down.
- Rate-limiting, to ensure that the rate at which the bad guy can make bad things happen can't make the system fail quickly. Recent DDoS attacks, such as the 400Gbps NTP Reflection attack on CloudFlare[6], show the importance of rate-limiting[7] services such as DNS and NTP.
- Lots of copies, so that the anti-entropy protocol can work by *sampling* the population of copies. Randomly sampling the peers makes it hard for the bad guy to know which peers are involved in which operations.

Now, our free, open source, peer-to-peer digital preservation system is in use at around 150 libraries worldwide, in about a dozen Private LOCKSS Networks (PLNs) preserving a variety of genres of content, and the original Global LOCKSS Network. One of the PLNs, the CLOCKSS Archive, a dark archive preserving the content of many academic publishers, was recently the fifth archive to be certified as a "Trustworthy Repository" by CRL[8], and the first to receive the highest possible score for its technical implementation. The program has been economically self-supporting for nearly 7 years[9] using the "RedHat" model of free software and paid support. In addition to our SOSP paper, the program has published research into many aspects of digital preservation.[10]

The P2P architecture of the LOCKSS system is unusual among digital preservation systems for a specific reason. The goal of the system was to preserve published information, which one has to assume is covered by copyright. One hour of a good copyright lawyer will buy, at current prices, about 12TB of disk, so the design is oriented to making efficient use of lawyers rather than disk.

2 Stuff is going to get lost

Since 2007 I've been using an example[11] of digital preservation in its most abstract form; a black box into which you put a Petabyte, and out of which a century later you take a Petabyte. Inside the box there can be as much redundancy as you want, on whatever media you choose, managed by whatever anti-entropy protocols you want. The design goal is a 50 % chance that every bit is the same coming out as going in. Consider every bit as like a radioactive atom, subject to a random process that flips it with a very low probability. You have just specified a half-life for the bits. It is about 60 million times the age of the universe. It isn't

[5] Maniatis, P.; Roussopoulos, M.; Giuli, T. J.; Rosenthal, D. S. F.; Baker, M.; Muliadi, Y.: Preserving Peer Replicas By Rate-Limited Sampled Voting. [Paper presented at the] 19th ACM Symposium on Operating Systems Principles, Bolton Landing, NY, October 2003, http://www.eecs.harvard.edu/~mema/publications/SOSP2003.pdf (last downloaded 2015-01-07).

[6] Prince, M.: Technical Details Behind a 400Gbps NTP Amplification DDoS Attack. In: CloudFlare blog from February 13, 2014, http://blog.cloudflare.com/technical-details-behind-a-400gbps-ntp-amplification-ddos-attack (last downloaded 2015-01-07).

[7] Vixie, P.: The edge of the Internet is an unruly place. In: acmqueue blog from February 4, 2014, https://queue.acm.org/detail.cfm?id=2578510 (last downloaded 2015-01-07).

[8] CRL, Certification Report on the CLOCKSS Archive, 2014. http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/clockss-report (last downloaded 2015-01-07).

[9] Rosenthal (footnote 2).

[10] LOCKSS Publications [list], http://www.lockss.org/news-media/publications/ (last downloaded 2015-01-07).

[11] Rosenthal, D. S. H.: "Petabyte for a Century" Goes Mainstream. In: DSHR's Blog from October 6, 2010, http://blog.dshr.org/2010/09/petabyte-for-century-goes-main-stream.html (last downloaded 2015-01-07).

feasible to benchmark a system to show that no process with a half-life *less* than 60 million times the age of the universe operates in it. Since at scale you are never going to *know* that your system is reliable enough, Murphy's law will guarantee that it isn't.

At scale, storing realistic amounts of data for human timescales is an unsolvable problem. Stuff is going to get lost. This shouldn't be a surprise, even in the days of paper stuff got lost. But the essential information needed to keep society running, to keep science progressing, to keep the populace entertained was stored very robustly, with many copies on durable, somewhat tamper-evident media in a fault-tolerant, P2P, geographically and administratively diverse system. This is no longer true. The Internet has, in the interest of reducing costs and speeding communication, removed the redundancy, the durability and the tamper-evidence from the system that stores society's critical data. It's now all on spinning rust, hopefully with at least one backup on rust-covered tape.

Recently, Berkeley researchers co-authored a paper reporting that:

> „[...] a rapid succession of coronal mass ejections [...] sent a pulse of magnetized plasma barreling into space and through Earth's orbit. Had the eruption come nine days earlier, when the ignition spot on the solar surface was aimed at Earth, it would have hit the planet, potentially wreaking havoc with the electrical grid, disabling satellites and GPS, and disrupting our increasingly electronic lives. [...] A study last year estimated that the cost of a solar storm like [this] could reach $2.6 trillion worldwide."[12]

Most of the information needed to recover from such an event exists only in digital form on magnetic media. Much of it probably exists only in "the cloud", happily immune to the electromagnetic effects of coronal mass ejections and easy to access after the power grid goes down.[13]

3 Why is stuff going to get lost?

One way to express the "what could possibly go wrong?" question is to ask "against what threats are you trying to preserve data?" The threat model of a digital preservation system is a very important aspect of the design which is, alas, only rarely documented. In 2005 we documented the LOCKSS threat model[14], observing that most discussion of digital preservation focused on these threats:

- Media failure
- Hardware failure
- Software failure
- Network failure
- Obsolescence
- Natural Disaster,

but that in the experience of operators of large data storage facilities the significant causes of data loss were quite different:

- Operator Error
- External Attack
- Insider Attack
- Economic Failure
- Organizational Failure

Unfortunately, we didn't consider coronal mass ejections or societal collapse from global warming.

The more we spend per byte, the safer the bytes are going to be, but this is subject to the Law of Diminishing Returns. Each successive nine of reliability is exponentially more expensive than the last. We don't have an unlimited budget, so we're going to have to trade off cost against the probability of data loss. To do this we need models to predict the cost of storing data using a given technology, and models to predict the probability of that technology losing data. I've worked on both kinds of model and can report that they're both extremely difficult.

Research, from among others Google[15], C-MU[16] and BackBlaze[17], shows that failure rates of storage media in service are much higher than the rates claimed by the

[12] Sanders, R.: Fierce Solar Magnetic Storm Barely Missed Earth in 2012. In: UC Berkely News Center from March 18, 2014, http://newscenter.berkeley.edu/2014/03/18/fierce-solar-magnetic-storm-barely-missed-earth-in-2012/ (last downloaded 2015-01-07).

[13] Rosenthal, D. S. H.: Coronal Mass Ejections. In: DSHR's Blog from July 25, 2014, http://blog.dshr.org/2014/07/coronal-mass-ejections.html (last downloaded 2015-01-07).

[14] Rosenthal, D. S. H.; Robertson, T. S.; Lipkis, T.; Reich, V.; Morabito, S.: Requirements for Digital Preservation Systems: A Bottom-Up Approach. In: D-Lib Magazine 11 (11) (2005).

[15] Pinheiro, E.; Weber, W.-D.; Barroso, L. A.: Failure Trends in a Large Disk Drive Population. In: Proceedings of the 5th USENIX Conference on File and Storage Technologies, San Jose, CA, February 13–16, 2007, https://www.usenix.org/legacy/events/fast07/tech/pinheiro.html (last downloaded 2015-01-07).

[16] Schroeder, B.; Gobson, G. A.: Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You? In: Proceedings of the 5th USENIX Conference on File and Storage Technologies, San Jose, CA, February 13–16, 2007, https://www.usenix.org/legacy/events/fast07/ (last downloaded 2015-01-07).

[17] Beach, B.: What Hard Drive Should I Buy? In: Backblaze Blog from January 21, 2014, http://blog.backblaze.com/2014/01/21/what-hard-drive-should-i-buy/ (last downloaded 2015-01-07).

manufacturer's specifications. For example, the Blu-Ray disks Facebook is experimenting with for cold storage[18] claim a 50-year data life. No-one has seen a 50-year-old DVD disk, so how do they know? The claims are based on a model of the failure mechanisms and data from accelerated life testing[19], in which batches of media are subjected to unrealistically high temperature and humidity. The model is used to extrapolate from these unrealistic conditions to the conditions to be encountered in service. But the conditions in service typically don't match those assumed by the models, and the models only capture some of the failure mechanisms.

These problems are much worse when we try to model not just failures of individual media, but of the entire storage system. Research has shown that media failures account for less than half the failures encountered in service[20]; other components of the system such as buses, controllers, power supplies, and so on contribute the other half. But even models that include these components exclude many of the threats we identified, from operator errors to coronal mass ejections.

Even more of a problem is that the threats, especially the low-probability ones, are highly correlated. Operators are highly likely to make errors when they are stressed coping with, say, an external attack.[21] The probability of economic failure is greatly increased by, say, insider abuse. Modeling these correlations is a nightmare.

It turns out that economics are by far the largest cause of data failing to reach future readers. At a Berkeley iSchool seminar entitled *The Half-Empty Archive*[22], I pulled together the various attempts to measure how much of the data that should be archived is actually being collected by archives, assessing that it was much less than half. No-one believes that archiving budgets are going to increase greatly. The loss rate from *can't afford to collect* will be at least 50 %, dwarfing all other causes.

4 Let's keep everything forever!

Digital preservation has three cost areas; ingest, preservation and dissemination. In the seminar I looked at the prospects for radical cost decreases in all three. Space here restricts me to storage, which is the main cost of preservation. The expectation is that, if not actually free, storage is so cheap we can afford to store everything forever.[23] This expectation comes from a third of a century of *Kryder's Law*[24], the analog of Moore's Law for disks. Kryder's Law predicted that the bit density on the platters of disk drives would more than double every 18 months, leading to a consistent 30–40 %/yr drop in cost per byte. Thus, long-term storage was effectively free. If you could afford to store something for a few years, you could afford to store it forever. The cost would have become negligible.

In the real world exponential growth can't continue forever[25]; it is always the first part of an S-curve. This graph, from Preeti Gupta at UCSC, shows that in 2010, even before the floods in Thailand doubled $/GB overnight, the Kryder curve was flattening. Currently, disk is about seven times as expensive as it would have been had the pre-2010 Kryder's Law continued. Industry projections are for 10–20 %/yr going forward – the red lines on the graph show that in 2020 disk is now expected to be 100–300 times more expensive than pre-2010 expectations.

Industry projections have a history of optimism, but if we believe that data grows at IDC's 60 %/yr[26], disk density

[18] Brodkin, J.: Facebook uses 10,000 Blu-ray Discs to Create Petabytes of "Cold Storage". In: ars technica, Technology Lab / Information Technology [blog] from January 29, 2014, http://arstechnica.com/information-technology/2014/01/facebook-uses-10000-blu-ray-discs-to-create-petabytes-of-cold-storage/ (last downloaded 2015-01-07).

[19] Williams, P.; Rosenthal, D. S. H.; Roussopoulos, M.; Georgis, S.: Predicting the Archival Life of Removable Hard Disk Drives. In: Proceedings of Imaging Science and Technology (ISandT) Archiving Conference, Bern, Switzerland, June 2008, http://www.eecs.harvard.edu/~mema/publications/removableDisks-2008.pdf (last downloaded 2015-01-07).

[20] Jiang, W.; Hu, C.; Zhou, Y.; Kanevsky, A.: Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. In: Proceedings of the 6th USENIX Conference on File and Storage Technologies, Napa Valley, CA, February 25–26, 2008, https://www.usenix.org/legacy/events/fast08/ (last downloaded 2015-01-07).

[21] Reason, J.: Human Error. Cambridge [England], New York, NY 1990.

[22] Rosenthal, D. S. H.: The Half-Empty Archive. In: DSHR's Blog from March 31, 2014, http://blog.dshr.org/2014/03/the-half-empty-archive.html (last downloaded 2015-01-07).

[23] Rosenthal, D. S. H.: Let's Just Keep Everything Forever In The Cloud. In: DSHR's Blog from May 14, 2012, http://blog.dshr.org/2012/05/lets-just-keep-everything-forever-in.html (last downloaded 2015-01-07).

[24] Walter, C.: Kryder's Law. In: Scientific American, August 2005, http://www.scientificamerican.com/article/kryders-law/ (last downloaded 2015-01-07).

[25] Rosenthal, D. S. H.: Paying For Long-Term Storage Revisited. In: DSHR's Blog from July 29, 2011, http://blog.dshr.org/2011/07/paying-for-long-term-storage-revisited.html (last downloaded 2015-01-07).

[26] IDC, The 2011 IDC Digital Universe Study. http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf (last downloaded 2015-02-02).

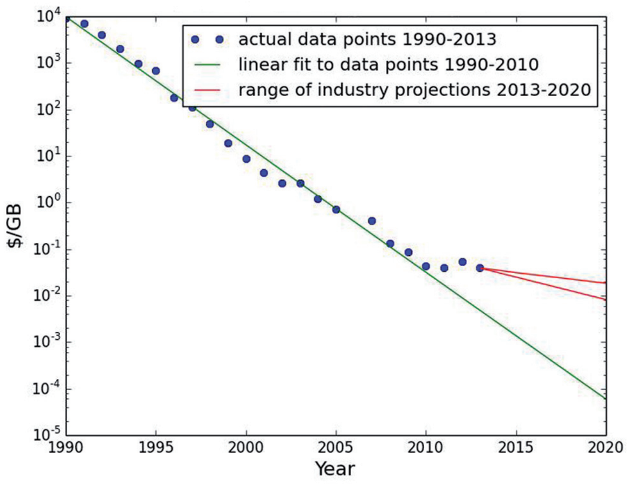

Figure 1: Disk cost-per-byte © P. Gupta

grows at IHS iSuppli's 20 %/yr[27], and IT budgets are essentially flat[28], the annual cost of storing a decade's accumulated data is 20 times the first year's cost. If at the start of the decade storage is 5 % of your budget, at the end it is more than 100 % of your budget.

We often[29] conflate Kryder's Law, which describes the increase in the areal density of bits on disk platters, with the cost of disk storage in $/GB, saying it roughly maps one-for-one into a decrease in the cost of disk drives.[30] Daniel Rosenthal has investigated the relationship between bits/in^2 and $/GB over the last couple of decades and concludes that about 3/4 of the decrease in $/GB can be attributed to the increase in bits/in^2. There are three possible causes for the remaining 1/4:

- **Economies of scale.** For most of the last two decades unit shipments of drives have been increasing, resulting in lower fixed costs per drive. Unfortunately, unit shipments are currently declining[31], so this effect has gone into reverse.
- **Manufacturing technology.** The technology to build drives has improved greatly over the last couple of decades, resulting in lower variable costs per drive. Unfortunately HAMR, the next generation of disk drive technology has proven to be extraordinarily hard to manufacture[32], so this effect has gone into reverse.
- **Vendor margins.** Over the last couple of decades disk drive manufacturing was a very competitive business, with numerous competing vendors. Unfortunately, the lack of competition and the floods have led to a major increase in margins[33], so this effect has gone into reverse.

These factors aren't enough to account for the slowing. A 2008 graph from Dave Anderson of Seagate shows how what looks like a smooth Kryder's Law curve is actually the superimposition of a series of S-curves, one for each successive technology. It shows Perpendicular Magnetic Recording (PMR) being replaced by Heat Assisted Magnetic Recording (HAMR) starting in 2009. No-one has yet shipped HAMR drives; the industry has resorted to stretching PMR by shingling (increasing the density) and helium (increasing the number of platters).

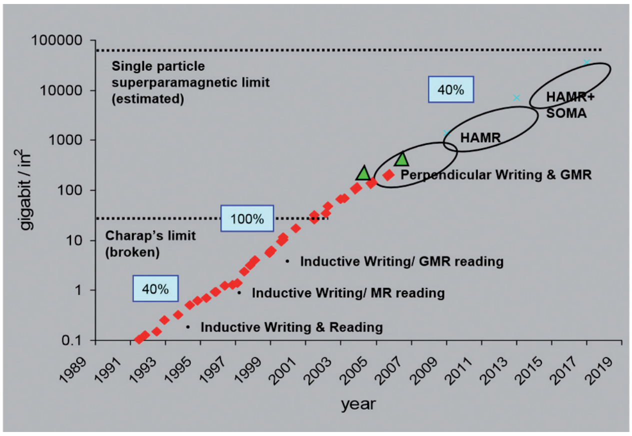

Figure 2: Area density growth[34]

[27] IHS iSuppli: HDD Areal Density Doubling in Five Years. In: Storage Newsletter May 24, 2012, http://www.storagenewsletter.com/news/marketreport/ihs-isuppli-storage-space (last downloaded 2015-01-07).

[28] http://computereconomics.com/ (last downloaded 2015-01-07).

[29] Rosenthal, D. S. H.: Bits per Square Inch vs. Dollars per GB. In: DSHR's Blog from November 12, 2012, http://blog.dshr.org/2012/11/bits-per-square-inch-vs-dollars-per-gb.html (last downloaded 2015-01-07).

[30] Kryder, M. H.; Kim, C. S.: After Hard Drives – What Comes Next? In: IEEE Transactions on Magnetics 45 (10) (2009) pp. 3406–3413.

[31] Rosenthal, D. S. H.: Peak Disk? In: DSHR's Blog from October 23, 2012, http://blog.dshr.org/2012/10/peak-disk.html (last downloaded 2015-01-07).

[32] Rosenthal, D. S. H.: Catching Up. In: DSHR's Blog from May 1, 2012, http://blog.dshr.org/2012/05/catching-up.html (last downloaded 2015-01-07).

[33] Rosenthal, D. S. H.: Storage Technology Update. In: DSHR's Blog from May 1, 2012, http://blog.dshr.org/2012/06/storage-technology-update.html (last downloaded 2015-01-07).

[34] From http://www.digitalpreservation.gov/meetings/documents/othermeetings/2-4_Anderson-seagate-v3_HDtrends.pdf (last downloaded 2015-02-11).

Each technology generation has to stay in the market long enough to earn a return on the cost of the transition from its predecessor. There are two problems:

- The return it needs to earn is, in effect, the margins the vendors enjoy. The higher the margins, the longer the technology needs to be in the market. Margins have increased.
- As technology advances, the easier problems get solved first. Each technology transition involves successively harder problems, costing more. The transition from PMR to HAMR has turned out to be vastly more expensive than the industry expected.

According to the 6-year-old graph, we should now be almost done with HAMR and starting the transition to Bit Patterned Media (BPM). This transition will be even more expensive and thus even more delayed than the PMR-HAMR transition. The projected 20 %/yr Kryder rate for disk is unlikely to be realized.

5 Alternatives to disk

About 70 % of the storage produced each year is disk[35], the rest being tape, optical and solid state. Tape has been the traditional medium for long-term storage. Its recording technology lags about 8 years behind disk; it is unlikely to run into the problems plaguing disk for some years. Its relative $/GB advantage over disk should grow in the medium term. But tape is losing ground in the market. Why is this?

Archived data[36] used to be rarely accessed; accesses other than integrity checks were sparse. Increasingly, as collections grow and data-mining tools become widely available, scholars want not to read individual documents, but to ask questions of the collection as a whole. Providing the compute power and I/O bandwidth to permit data-mining of collections is much more expensive than simply providing occasional sparse read access. Some idea of how much can be gained by comparing the Amazon's S3, designed for data-mining type access patterns, with Glacier, designed for traditional archival access. Amazon's S3 is currently at least 2.5 times as expensive[37]; until recently it was 5.5 times[38]. Thus future archives will need to keep at least one copy of their content on low-latency, high-bandwidth storage, not tape.

Flash memory's advantages, including low power, physical robustness and low access latency have overcome its higher cost per byte in many markets, such as tablets and servers. But there is no possibility of flash replacing disk in the bulk storage market; that would involve trebling the number of flash fabs. Even if we ignore the lead time to build the new fabs, the investment to do so would not pay dividends. Shrinking flash cells much further will impair their ability to store data. Increasing levels, stacking cells in 3D and increasingly desperate signal processing in the flash controller will keep density going for a little while, but not long enough to pay back the investment in the fabs.

There are many technologies vying to be the successor to flash and if the semiconductor industry keeps on its road-map they will keep scaling beyond the end of flash. They all have significant advantages over flash, in particular being byte- rather than block-addressable. But analysis by Mark Kryder and Chang Soo Kim at Carnegie-Mellon is not encouraging about the prospects for either flash or the competing solid state technologies beyond the end of the decade.[39]

Every few months there is another press release announcing that some new, quasi-immortal medium[40] such as stone DVDs has solved the problem of long-term storage. But the problem stays resolutely unsolved. Very long-lived media are inherently more expensive, and are a niche market, so they lack economies of scale. Seagate could easily make disks with archival life[41], but a study of the market for them revealed that no-one would pay the relatively small additional cost. The fundamental problem

[35] Mellor, C.: WD Bigshots Spin Superfast Disk Roadmap. In: The Register [blog] from May 9, 2012, http://www.theregister.co.uk/2012/05/09/wd_disk_tech_views/ (last downloaded 2015-01-07).

[36] Adams, I. F.; Storer, M. W.; Miller, E. L.: Analysis of Workload Behavior in Scientific and Historical Long-term Data Repositories. In: ACM Transactions on Storage 8 (2) (2012) Article 6. DOI:http://dx.doi.org/10.1145/2180905.2180907 (last downloaded 2015-01-07).

[37] Amazon S3 Pricing, https://aws.amazon.com/s3/pricing/ (last downloaded 2015-01-07).

[38] Clark, J.: Google Slashes Cloud Storage to $0.026 per GB. Your move, Amazon. In: The Register [blog] from March 25, 2014, http://www.theregister.co.uk/2014/03/25/google_price_slash/ (last downloaded 2015-01-07).

[39] Kryder, M. H., and C. S. Kim (footnote 30).

[40] Rosenthal, D. S. H.: Immortal Media. In: DSHR's Blog from July 16, 2013, http://blog.dshr.org/2013/07/immortal-media.html (last downloaded 2015-01-07).

[41] Anderson, D.: Archive Drive Study, http://www.digitalpreservation.gov/meetings/documents/othermeetings/5-4_Anderson-seagate-v3_archive_study.pdf (last downloaded 2015-01-07).

is that long-lived media only make sense at very low Kryder rates. Even if the rate is only 10 %/yr, after 10 years you could store the same data in 1/3 the space. Since space in the data center, or even at Iron Mountain, isn't free this is a powerful incentive to move old media out. If you believe that Kryder rates will get back to 30 %/yr, after a decade you could store 30 times as much data in the same space.

The reason that the idea of long-lived media is so attractive is that it suggests that you can be lazy and design a system that ignores the possibility of failures. You can't:

- Media failures are only one of many threats to stored data[42], but they are the only one long-lived media address.
- Long media life does not imply that the media are more reliable, only that their reliability decreases with time more slowly. As we have seen, current media are many orders of magnitude too unreliable for the task ahead.

Even if you could ignore failures, it wouldn't make economic sense. As Brian Wilson, CTO of BackBlaze points out, in their long-term storage environment:

> "Double the reliability is only worth 1/10th of 1 percent cost increase. [...] Replacing one drive takes about 15 minutes of work. If we have 30,000 drives and 2 percent fail, it takes 150 hours to replace those. In other words, one employee for one month of 8 hour days. Getting the failure rate down to 1 percent means you save 2 weeks of employee salary – maybe $5,000 total? The 30,000 drives costs you $4m. The $5k/$4m means the Hitachis are worth 1/10th of 1 per cent higher cost to us."[43]

Note that this analysis assumes that the drives fail under warranty. One thing the drive vendors did to improve their margins after the floods was to reduce the length of warranties.

6 Does Kryder's Law slowing matter?

Figures from SDSC[44] suggest that media cost is about 1/3 of the lifecycle cost of storage, although figures from BackBlaze[45] suggest a much higher proportion. As a rule of thumb, the research into digital preservation costs suggests that ingesting the content costs about 1/2 the total lifecycle costs, preserving it costs about 1/3 and disseminating it costs about 1/6. So why are we worrying about a slowing of the decrease in 1/9 of the total cost?

Different technologies with different media service lives involve spending different amounts of money at different times during the life of the data. To make apples-to-apples comparisons we need to use the equivalent of Discounted Cash Flow[46] to compute the *endowment* needed for the data. This is the capital sum which, deposited with the data and invested at prevailing interest rates, would be sufficient to cover all the expenditures needed to store the data for its life.

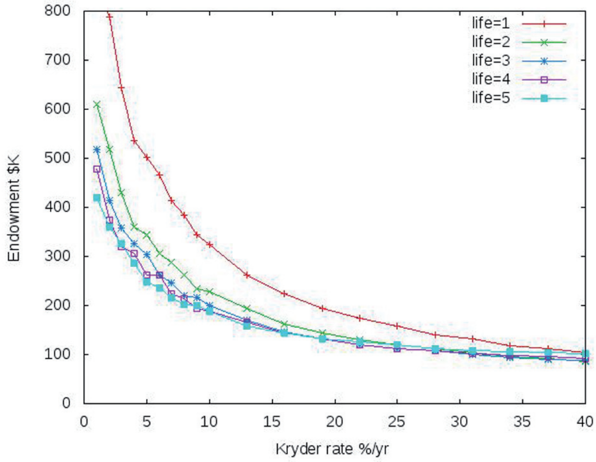

Figure 3: Kryder rate © D. S. H. Rosenthal

[42] Rosenthal et al. (footnote 14).

[43] Olds, D.: How NOT to Evaluate Hard Disk Reliability: Backblaze vs World+dog. In: The Register [blog] from February 17, 2014, http://www.theregister.co.uk/2014/02/17/backblaze_how_not_to_evaluate_disk_reliability/ (last downloaded 2015-01-07).

[44] Moore, R. L.; D'Aoust, J.; McDonald, R. H.; Minor, D.: Disk and Tape Storage Cost Models. San Diego, CA [2007], http://users.sdsc.edu/~mcdonald/content/papers/dt_cost.pdf (last downloaded 2015-01-07).

[45] Nufire, T.: Petabytes on a Budget v2.0. Revealing More Secrets. In: Backblaze Blog from July 20, 2011, http://blog.backblaze.com/2011/07/20/petabytes-on-a-budget-v2-0revealing-more-secrets/ (last downloaded 2015-01-07).

[46] Rosenthal, D. S. H.: Paying For Long-Term Storage Revisited. In: DSHR's Blog from July 29, 2011, http://blog.dshr.org/2011/07/paying-for-long-term-storage-revisited.html (last downloaded 2015-01-07).

We built an economic model of the cost of long-term storage.[47] Here it plots the endowment needed for 3 replicas of a 117TB dataset to have a 98 % chance of not running out of money over 100 years, against the Kryder rate, using costs from BackBlaze. Each line represents a policy of keeping the drives for 1, 2 ... 5 years before replacing them.

In the past, with Kryder rates in the 30–40 % range, we were in the flatter part of the graph where the precise Kryder rate wasn't that important in predicting the long-term cost. As Kryder rates decrease, we move into the steep part of the graph, which has two effects. The endowment needed:

- Increases sharply.
- Becomes harder to predict, because it depends strongly on the precise Kryder rate.

The reason to worry is that the cost of storing data for the long term depends strongly on the Kryder rate if it falls much below 20 %, which it has. Everyone's storage expectations, and budgets, are based on their pre-2010 experience, and on a belief that the effect of the floods was a one-off glitch and the industry will quickly get back to historic Kryder rates. It wasn't, and they won't.

7 Does losing stuff matter?

Consider two storage systems with the same budget over a decade, one with a loss rate of zero, the other half as expensive per byte but which loses 1 % of its bytes each year. Clearly, you would say the cheaper system has an unacceptable loss rate. However, each year the cheaper system stores twice as much and loses 1 % of its accumulated content. At the end of the decade the cheaper system has preserved 1.89 times as much content at the same cost. After 30 years it has preserved more than 5 times as much at the same cost.

Adding each successive nine of reliability gets exponentially more expensive. How many nines do we really need? Is losing a small proportion of a large dataset really a problem? The canonical example of this is the Internet Archive's web collection. Ingest by crawling the Web is a lossy process. Their storage system loses a tiny fraction of its content every year. Access via the Wayback Machine is not completely reliable. Yet as I write this, for US users archive.org is the 153rd most visited site[48], whereas loc.gov is the 1231st.[49]

Why is this? Because the collection was always a series of *samples* of the Web, the losses merely add a small amount of random noise to the samples. But the samples are so huge that this noise is insignificant. This isn't something about the Internet Archive, it is something about very large collections. In the real world they always have noise; questions asked of them are always statistical in nature. The benefit of doubling the size of the sample vastly outweighs the cost of a small amount of added noise. In this case more is better.

8 Summing up

Cost is the biggest barrier to preservation. Cost expectations were formed during three decades of extremely rapid storage cost decrease. They remain unchanged despite four years of no decrease. Industry analysts are projecting no more than 20 %/yr rates for the rest of the decade. Technological and market forces make it likely that, as usual, they are being optimistic. Lower Kryder rates greatly increase the cost of long-term storage, its proportion of preservation costs, and the uncertainty in estimating them.

In the short term, the inertia of manufacturing investment means that things aren't going to change much. Bulk data is going to be on disk, it can't compete with other uses for the higher-value space on flash. The reason disks have a 5-year service life isn't an accident of technology; they are *engineered* to last that long because, at a 40 %/yr Kryder rate, it is uneconomic to keep the drive for longer than 5 years. After 5 years the data will take up about 8 % of the drive's replacement. At lower Kryder rates the media will be in service longer. That means that running cost will be a larger proportion of the total cost. It will be worthwhile to spend more on purchasing the media to spend less on running them.[50] This is particularly true since bulk storage media are no longer a consumer product; businesses are better placed to make this trade-off. But they

[47] Rosenthal, D. S. H.: Talk at IDCC2013. In: DSHR's Blog from January 22, 2013, http://blog.dshr.org/2013/01/talk-at-idcc2013.html (last downloaded 2015-01-07).

[48] http://www.alexa.com/siteinfo/archive.org (last downloaded 2015-01-07).

[49] http://www.alexa.com/siteinfo/loc.gov (last downloaded 2015-01-07).

[50] Adams, I.; Miller, E. L.; Rosenthal, D. S. H.: Using Storage Class Memory for Archives with DAWN, a Durable Array of Wimpy Nodes. Technical Report UCSC-SSRC-11–07. Santa Cruz, CA 2011, http://www.ssrc.ucsc.edu/Papers/ssrctr-11-07.pdf (last downloaded 2015-01-07).

may not do so (see Haldane and Davies[51], and Farmer and Geanakoplos[52]).

The idea that archived data can live on long-latency, low-bandwidth media is no longer the case. Future archival storage architectures must deliver adequate performance to sustain data-mining as well as low cost. Bundling computation into the storage medium is the way to do this.

David S. H. Rosenthal
LOCKSS Program
Stanford University Libraries
Green Library
557 Escondido Mall
Stanford
CA 94305-6062
USA
dshr@stanford.edu

[51] Haldane, A. G.; Davies, R.: The Short Long. [Speech at the] 29th Société Universitaire Européene de Recherches Financières Colloquium “New Paradigms in Money and Finance?”, Brussels May 2011, http://www.bankofengland.co.uk/publications/Documents/speeches/2011/speech495.pdf (last downloaded 2015-01-07).

[52] Farmer, J. D.; Geanakoplos, J.: Hyperbolic Discounting is Rational. Valuing the Far Future With Uncertain Discount Rates. New Haven, Conn. 2009 (Cowles Foundation Discussion Paper 1719), http://cowles.econ.yale.edu/P/cd/d17a/d1719.pdf (last downloaded 2015-01-07).