Selecting efficient and reliable preservation strategies:

modeling long-term information integrity using large-scale hierarchical discrete event simulation

Prepared for
IDCC 20200
Dublin

Micah Altman
MIT Libraries

Richard Landau
Program on Information Science

< http://tiny.cc/IDCC2020Altman >

Related Work

ARXIV Preprint:

[1912.07908] Selecting efficient and reliable preservation strategies: modeling long-term information integrity using large-scale hierarchical discrete event simulation

Abstract

This article addresses the problem of formulating efficient and reliable operational preservation policies that ensure bit-level information integrity over long periods, and in the presence of a diverse range of real-world technical, legal, organizational, and economic threats. We develop a systematic, quantitative prediction framework that combines formal modeling, discrete-event-based simulation, hierarchical modeling, and then use empirically calibrated sensitivity analysis to identify effective strategies.

Specifically, the framework formally defines an objective function for preservation that maps a set of preservation policies and a risk profile to a set of preservation costs, and an expected collection loss distribution. In this framework, a curator's objective is to select optimal policies that minimize expected loss subject to budget constraints. To estimate preservation loss under different policy conditions optimal policies, we develop a statistical hierarchical risk model that includes four sources of risk: the storage hardware; the physical environment; the curating institution; and the global environment. We then employ a general discrete event-based simulation framework to evaluate the expected loss and the cost of employing varying preservation strategies under specific parameterization of risks.

The framework offers flexibility for the modeling of a wide range of preservation policies and threats. Since this framework is open source and easily deployed in a cloud computing environment, it can be used to produce analysis based on independent estimates of scenario-specific costs, reliability, and risks.

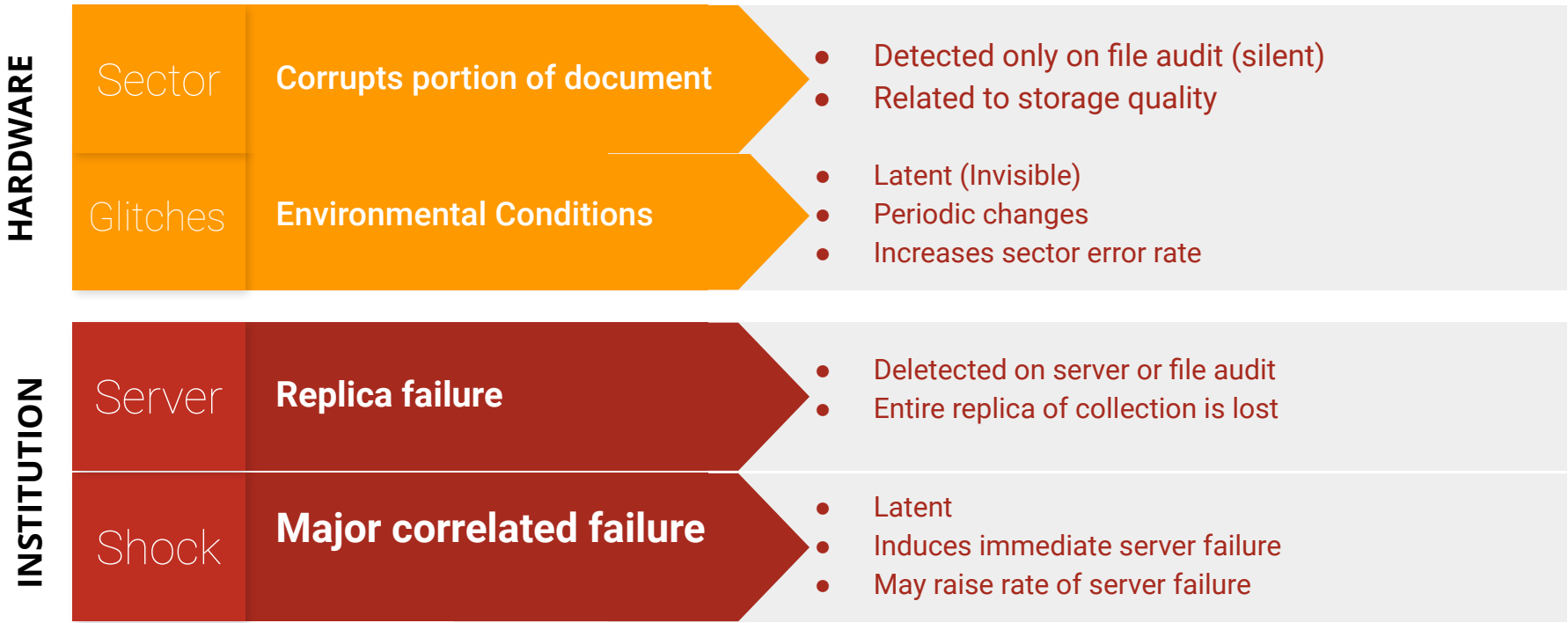
We present results summarizing hundreds of thousands of simulations using this framework. This analysis points to a number of robust and broadly applicable preservation strategies, provides novel insights into specific preservation tactics, and provides evidence that challenges received wisdom.

Shifting Economics of Digital Information

Going digital changes economics of long term access

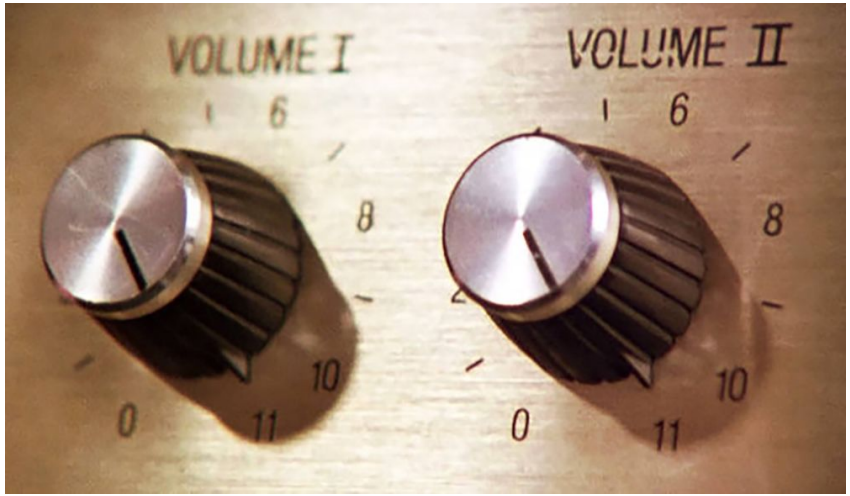
- Computation is cheap
- Replication is cheap
- Conservation
(of media, hardware)
is expensive

Multi-Level Threat Modeling



Core Preservation Actions

- Replication
- Auditing
- Repair
- Transformations
  - Compression
  - Encryption
  - Reformatting



Cost Modeling

Cost(C,S)= f(storage(C,S), communications(C,S), Replicas(S))

Simplifications:

- Each separate replication imposes a fixed cost
- Storage cost is linear in (compressed) collection size
- Communication is linear in collection size; audit frequency
- Other computation costs are negligible

Characterizing Preservation as Optimization

Given

- A *collection* (**C**), of documents ={**D1**..**DN**};
- A budget (**B**)
- Distribution of threats **P()**

Choose

A preservation strategy (**S**) =
{Copies, AuditMethod,
RepairFrequency, FileTransformation}

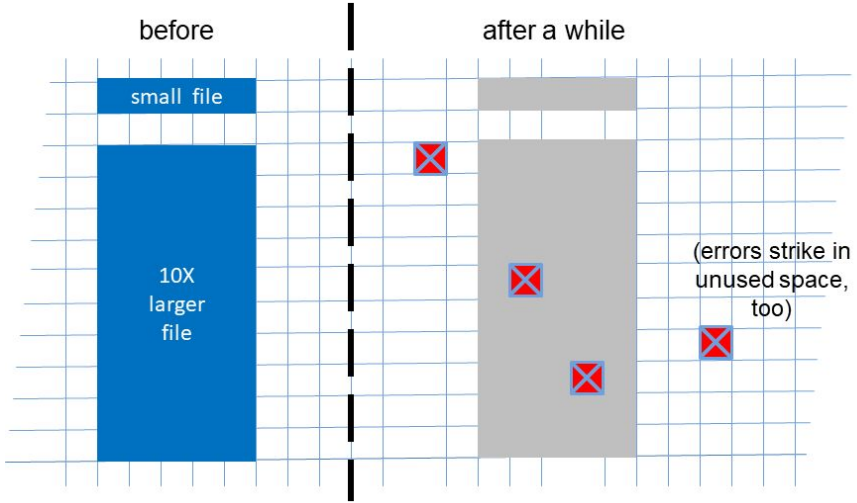
Optimize

Choose the optimal strategy, **S*,** to minimize collection loss, within the budget

$\min_{S* \ni S} E(Loss(C,S*)) \mid Cost(C,S*) \leq \mathbf{B}$

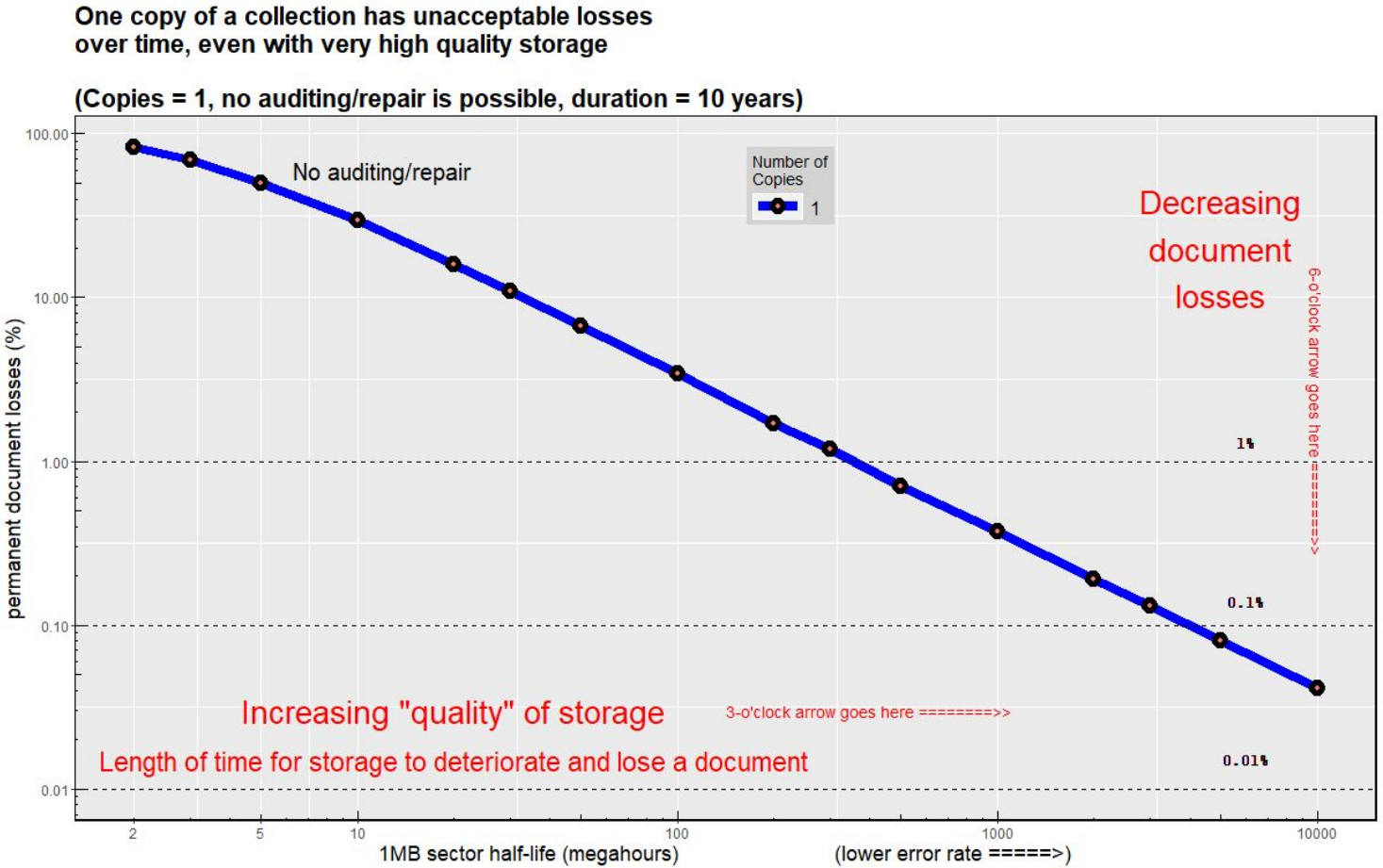
Protecting Against Hardware Errors

Modeling Damage to Files

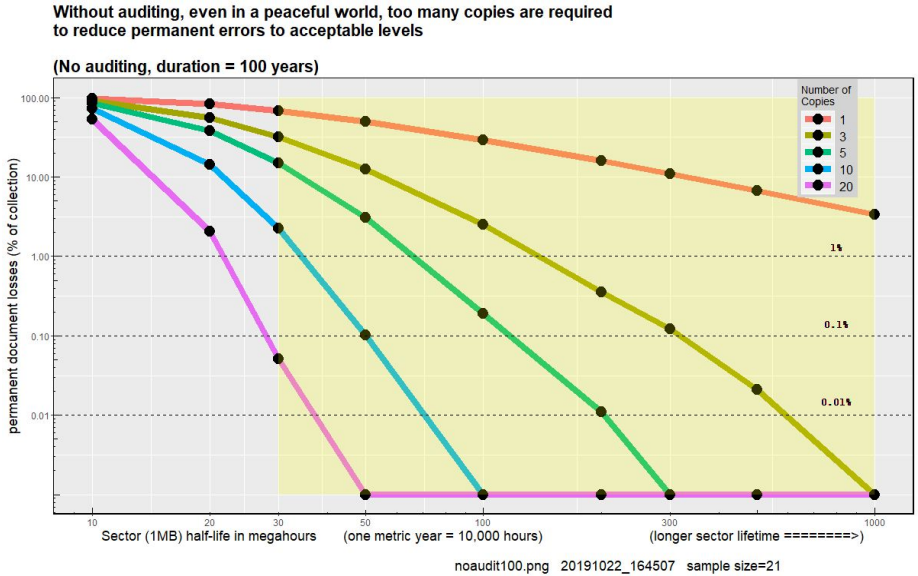


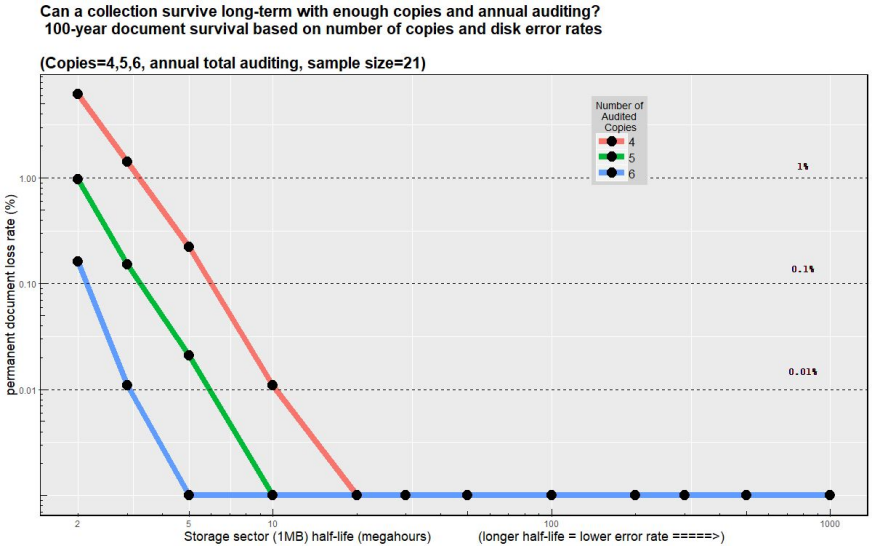
Layer	Role	Visibility	Distribution	Lower Frequency	Higher Frequency (lower severity)
Storage Hardware (Sector)	Causes sector error / single document loss	Silent.	Poisson event	Controller failure	Media corruption.
Local environment (Glitch)	Increases rate of storage error	Invisible.	Poisson event of some duration	HVAC failure	Power spikes

The Best Disk Hardware is Not Enough -- Make Copies



5 Copies + Systematic Annual Auditing is Sufficient Protection from Hardware Errors



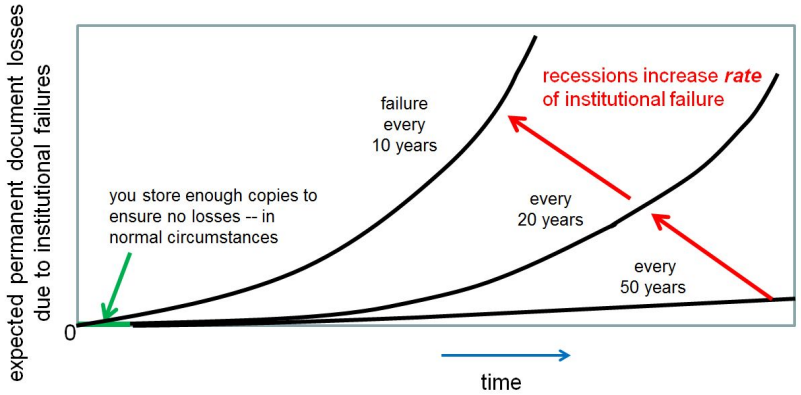


Without auditing 20 copies are necessary to prevent loss over a century

With simple annual auditing 5 Copies are Sufficient

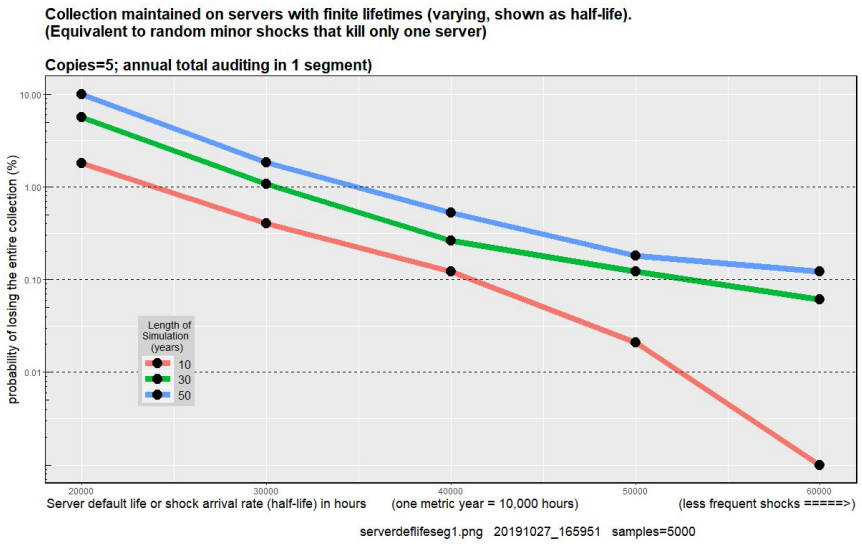
Institutional Level Threats

Modeling Institution-Level Failures

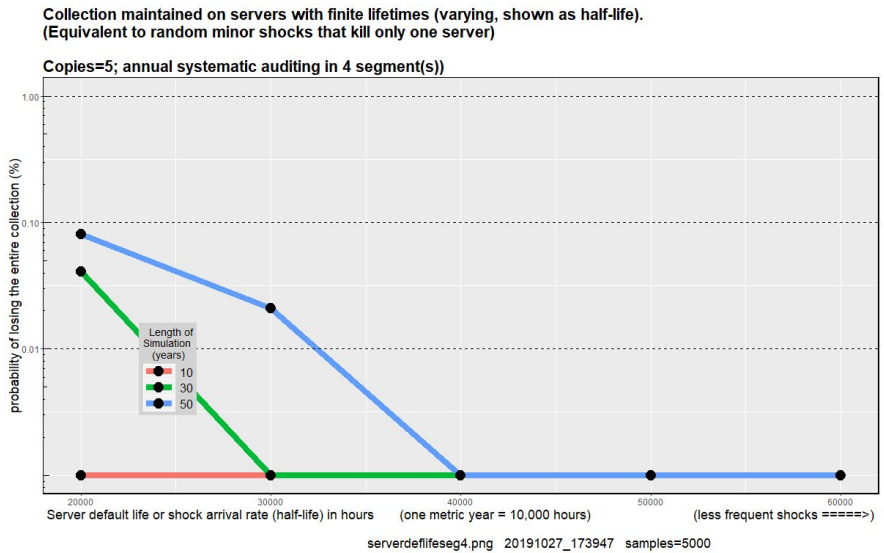


Layer	Role	Visibility	Distribution	Lower Frequency	Higher Frequency (lower severity)
Institution (Server Failure)	Causes loss of a single copy of a collection	Silent.	Exponential Lifetime	Ransomware Business failure	Curator error. Billing error
Macro Environment (Major Shock)	Increases rate of server failure	Invisible.	Poisson duration	Corporate Mergers	Recession
	Immediate loss of multiple servers	Silent or visible	Poisson event	Government Suppression	Regional war

Annual Auditing Does Not Protect Against Server Failure

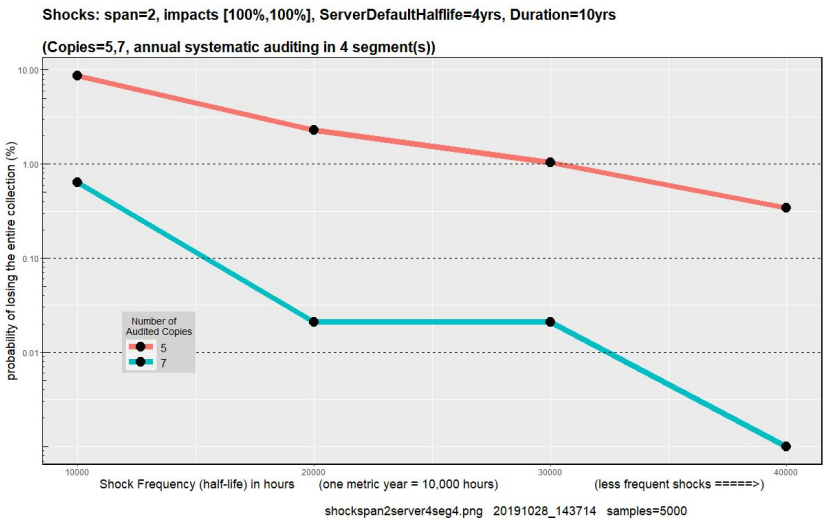


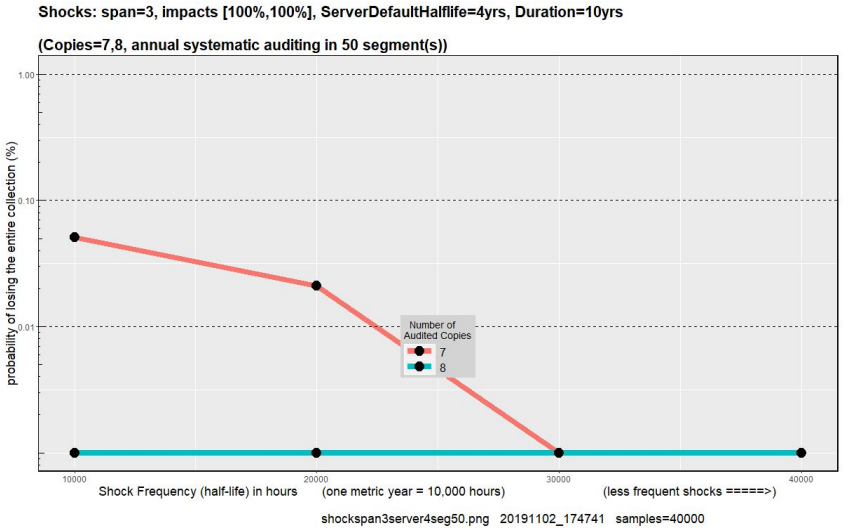
Annual Audits
Significant collection Loss over Long Term



Dividing Audit into Quarterly Segments Controls Risk

Protection for Major Recessions and Minor Wars





Weekly Server Auditing Protects Against Triple Simultaneous Server Failures

Managing Format Transformations

Managing File Encryption with Key Replication

Challenge

- Entire collection is encrypted, using a set of *E* encryption keys
- If all keys are destroyed, collection is lost

Modeling Risk

- Given key size, risk of loss (not corruption) dominates
- Model key failure as 'server' failure
- Audit action:
  - Challenge key-holder to prove it can decrypt

Results

➔ Replicate encryption keys (or partial shared secrets) across independent holders

➔ Shock size and frequency are driving factors

➔ Shock-resistant auditing strategy is sufficient

Managing Format Failure with Reader Verification

Challenge

- Documents are encoded
  - *K* formats in collection
- If format cannot be interpreted
  - Collection is loss

Modeling Risk

- Model format failure as ‘server’ failure
  - Each format *F* is maintained by *S* servers
  - Each S holds an executable reader that can read documents in that format
- Audit action:
  - Challenge server to prove it can read file

Results

➔ Migrate formats when number of functioning readers is below replication threshold

➔ Shock size and frequency are driving factors

➔ Shock-resistant auditing strategy is sufficient

Compress or Repair?

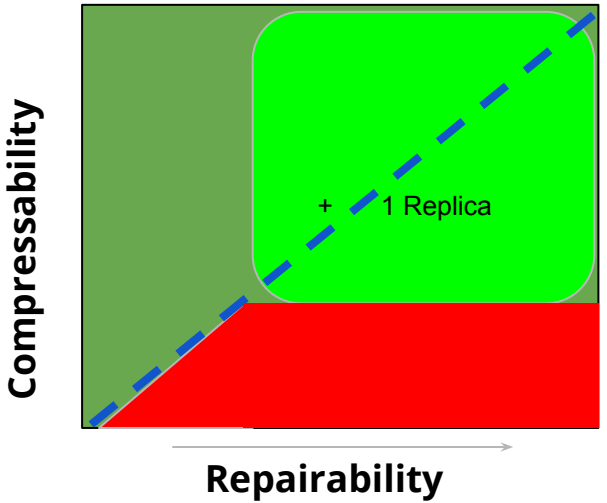
Benefits of compression

- Smaller document → reduce risk from hardware errors
- Smaller collection → more replicas can be purchased and audited for fixed cost

Risks of compression

- Increased fragility → single error destroys document
- Compression format must be managed for format obsolescence

Modeling



- Treat large repairable documents as collection of **R** smaller non-repairable documents
- Estimate using compression ratios for most common compression formats

Result: Compress

Recommendations

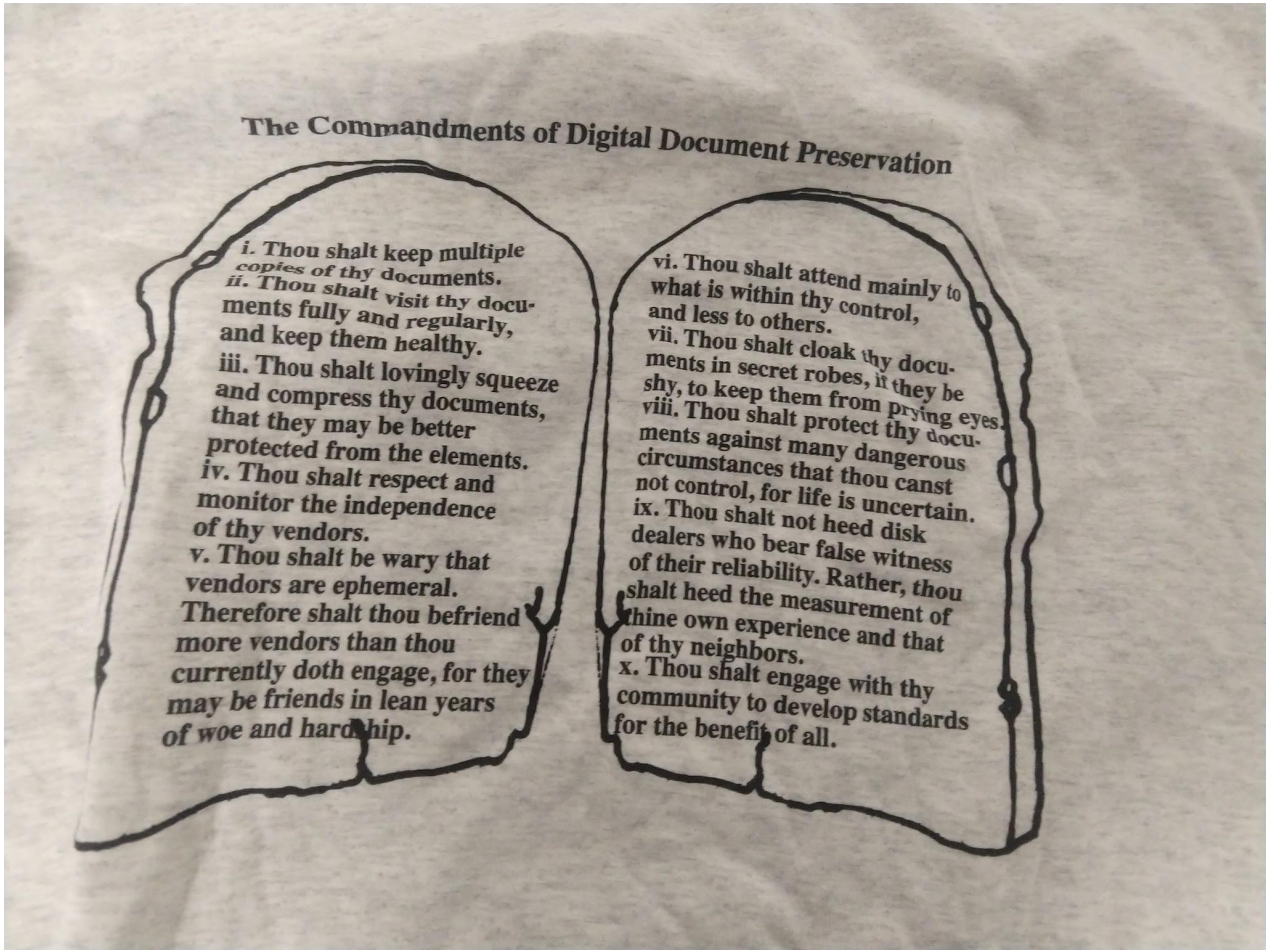
Bottom Line

for Memory Institutions

- Replicate
- Don’t fear the cloud
- Diversify across institutions
- Audit regularly and completely
- Audit storage, formats, secrets
- Compress

for Vendors

- Forget 11 nines … reveal replication strategy
- Collect and share loss rates
- Support auditing primitives
- Disclose institutional dependencies


The Commandments of Digital Document Preservation
i. Thou shalt keep multiple copies of thy documents.
ii. Thou shalt visit thy documents fully and regularly, and keep them healthy.
iii. Thou shalt lovingly squeeze and compress thy documents, that they may be better protected from the elements.
iv. Thou shalt respect and monitor the independence of thy vendors.
v. Thou shalt be wary that vendors are ephemeral. Therefore shalt thou befriend more vendors than thou currently doth engage, for they may be friends in lean years of woe and hardship.
vi. Thou shalt attend mainly to what is within thy control, and less to others.
vii. Thou shalt cloak thy documents in secret robes, if they be shy, to keep them from prying eyes.
viii. Thou shalt protect thy documents against many dangerous circumstances that thou canst not control, for life is uncertain.
ix. Thou shalt not heed disk dealers who bear false witness of their reliability. Rather, thou shalt heed the measurement of thine own experience and that of thy neighbors.
x. Thou shalt engage with thy community to develop standards for the benefit of all.