

Information Integrity Over the Long Term



Micah Altman

MIT Libraries

Richard Landau

Program on Information Science



Related Work

Draft - for internal comment:

<https://github.com/MIT-Informatics/PreservationSimulation>

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

The Tools of Preservation

- Replication
- Auditing
- Repair
- Compression



Characterizing Preservation as Optimization

Given

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

Choose

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

Optimize

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

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

Cost Modeling


$$\text{Cost}(C,S)=f(\text{storage}(C,S), \text{communications}(C,S), \text{Replicas}(S))$$

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

$$\rightarrow \mathbf{Cost(C,S)} = B1 * \text{Replicas} + \\ B2 * \text{AuditFrequency} * \text{Size}(C) + \\ B3 * \text{Size}(C) * \text{Replicas} * \text{CompressionFractor}(S)$$

Loss Modeling

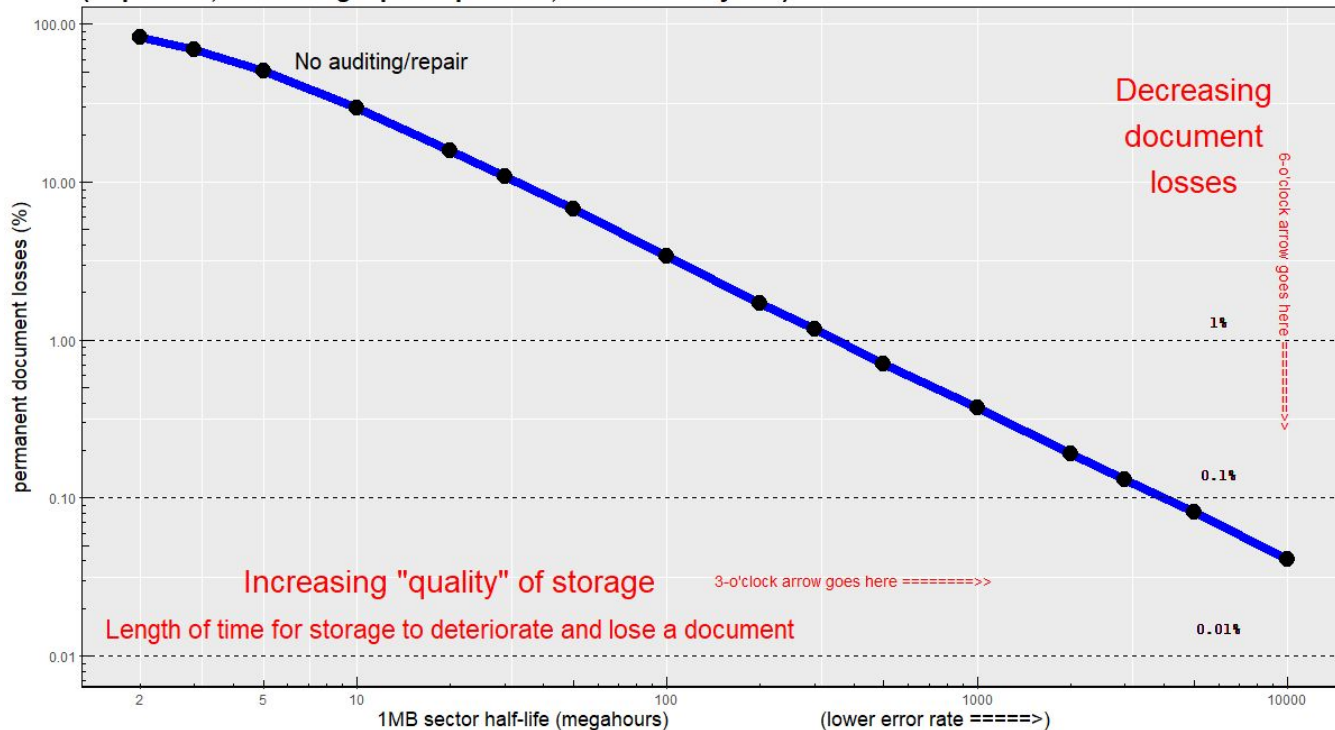
Sector	Corrupts portion of document	<ul style="list-style-type: none">•  Detected on audit (silent)• Exponentially distributed• Related to storage quality
Glitches	Environmental Conditions	<ul style="list-style-type: none">• Periodic changes• Increases sector error rate• Never directly observable (latent)
Server	Replica failure	<ul style="list-style-type: none">• Entire replica of collection is lost• Exponentially distributed
Shock	Major correlated failure	<ul style="list-style-type: none">• Induces immediate server failure• May raise rate of server failure

The Big Things

One Copy is Not Enough -- Even if Sector Error is Low

One copy of a collection has unacceptable losses over time, even with very high quality storage

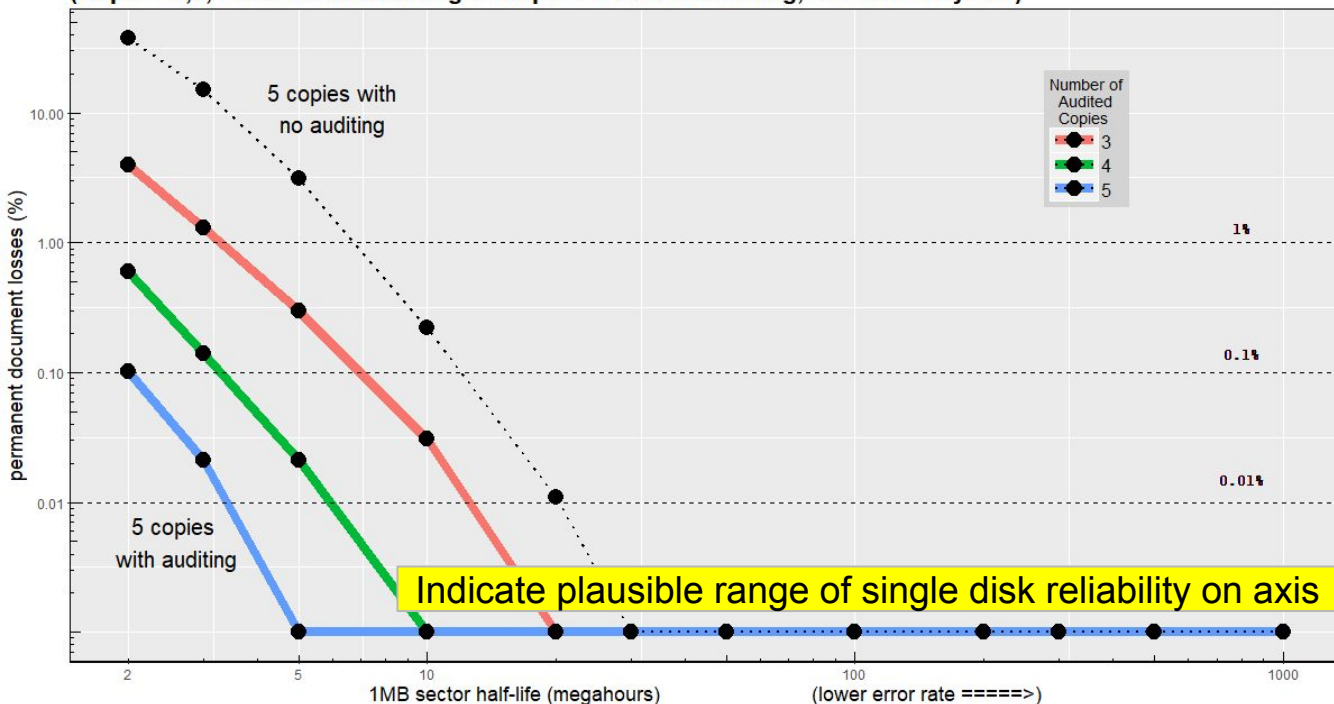
(Copies = 1, no auditing/repair is possible, duration = 10 years)



Some Copies + Auditing is better than Many Copies

With regular auditing, only a few copies are required to minimize losses over a wide range.
Failure to audit the collection is worse than keeping only a small number of audited copies

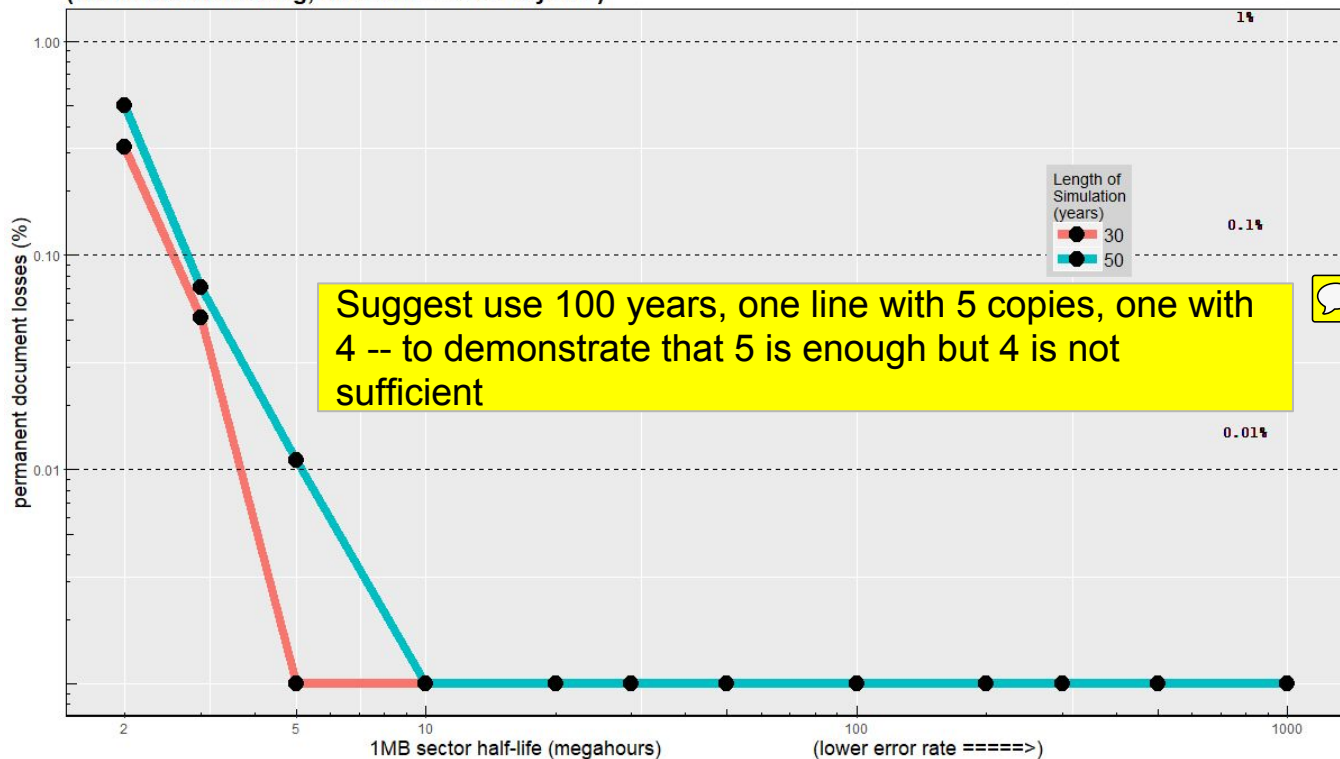
(Copies=3,4,5 with annual auditing vs copies=5 with no auditing, duration=10 years)



Five Copies (+ auditing) protects against low-level errors... Forever

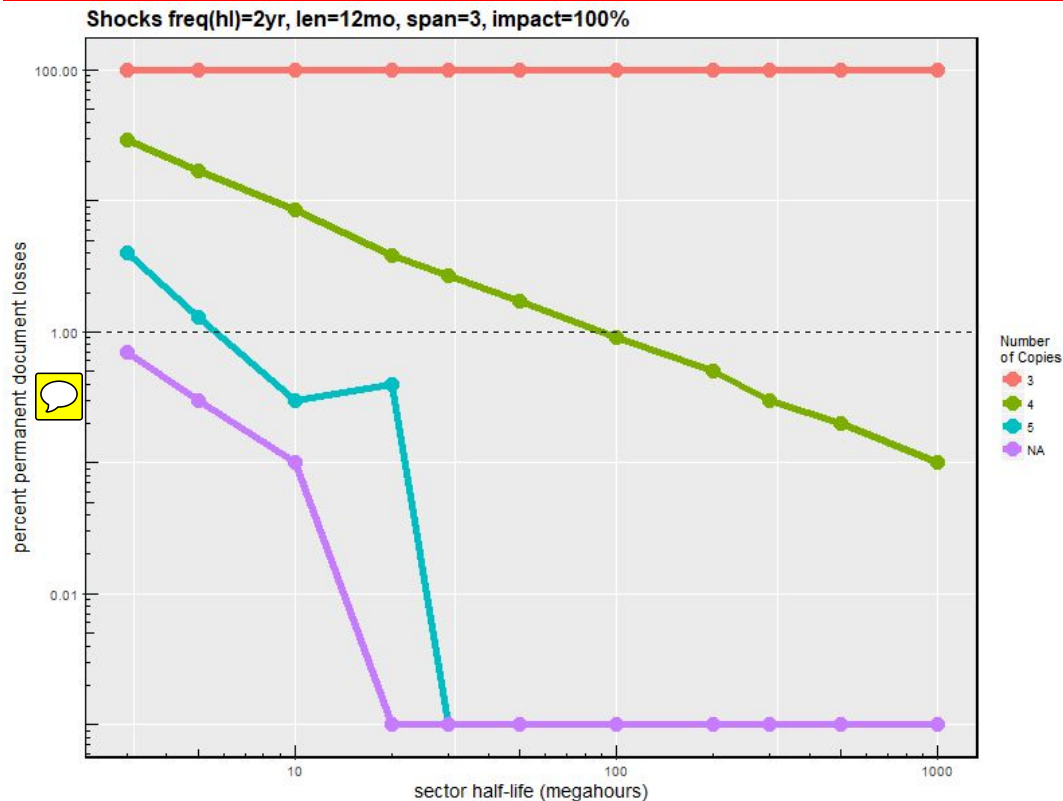
With moderate auditing, in a peaceful world, five copies are nearly immortal

(Annual total auditing, duration = 30 & 50 years)



(With enough copies...)

Sector error doesn't matter, server lifetime does



Shocks are Everywhere...

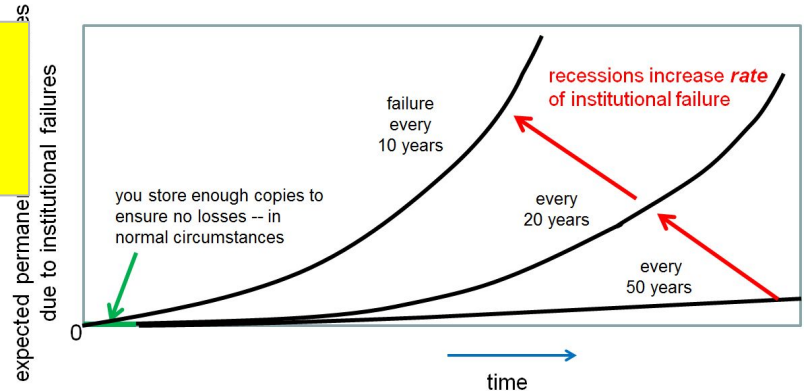
Single server failure?

Repression, Encryption
Key Loss, Financial

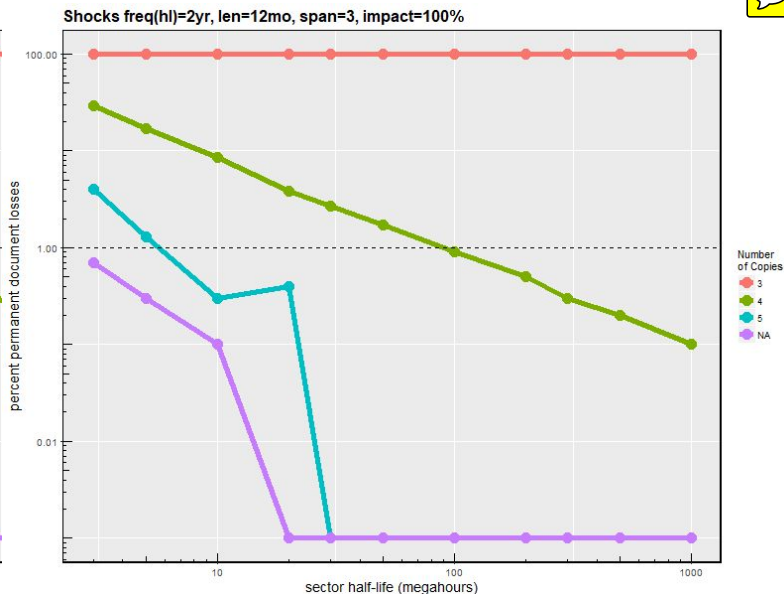
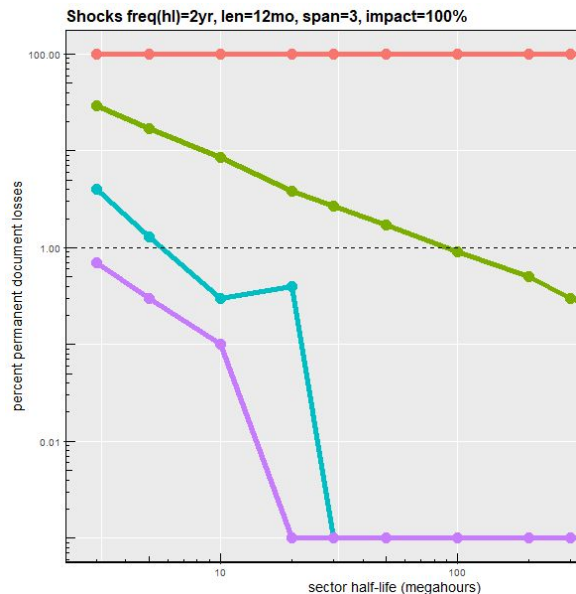
Companion abstract figure showing sudden
collapse?



Recession



Shocks matter -- even for long-lived servers



Seven (?) diversified copies will survive a major disaster or minor war



Suggest: fixed number of year 20?; Expected server lifetime of 5 years; Lines for 5,6,7 servers. X axis is increasing shock frequency for a major shock

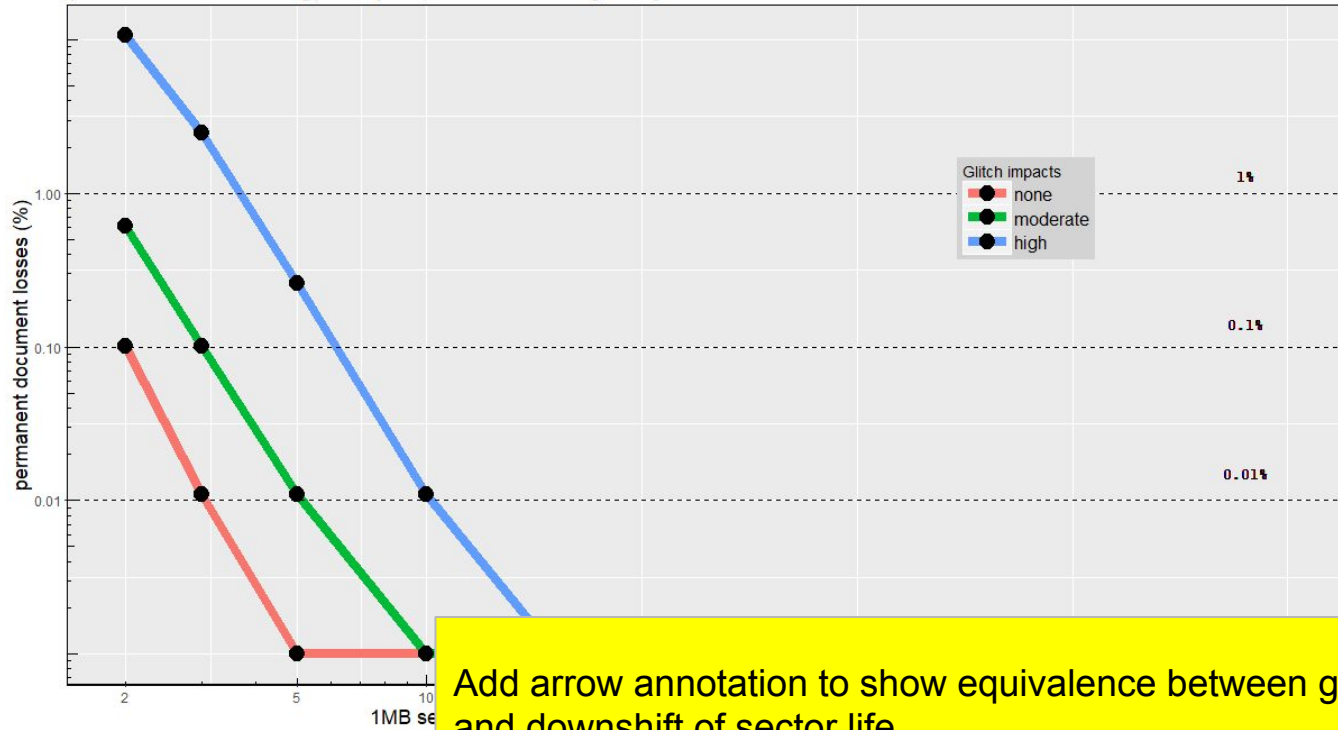
Complications (Do's and Don'ts)

Don't worry about glitches

Occasional temporary glitches increase the server error rate for some period, but otherwise are not substantially different from normal operation

Five copies is enough

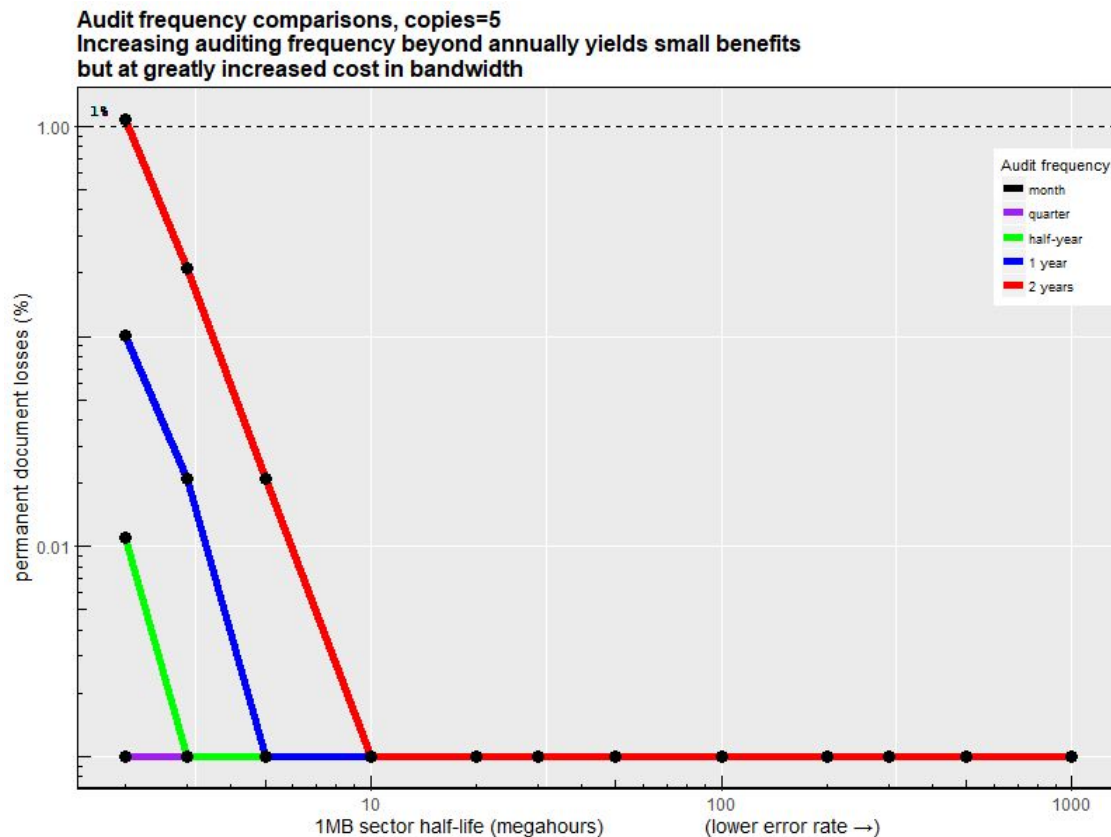
(Total annual auditing, 5 copies, duration = 10 years)



Add arrow annotation to show equivalence between glitch and downshift of sector life

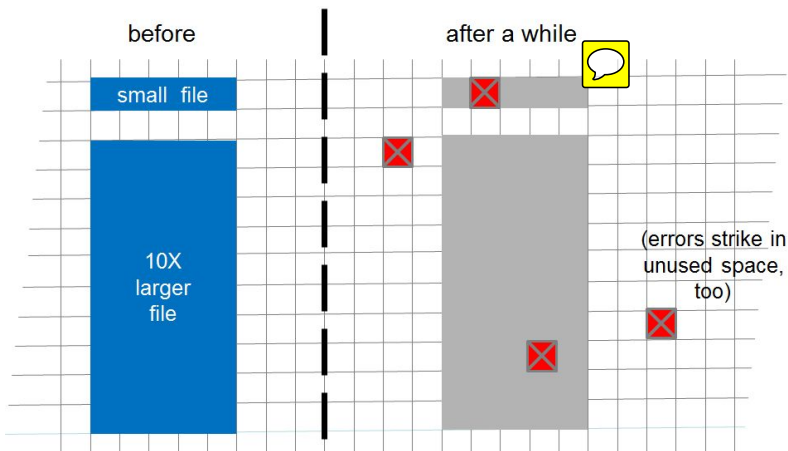
DON'T Worry about auditing frequency

-- Annually is Enough

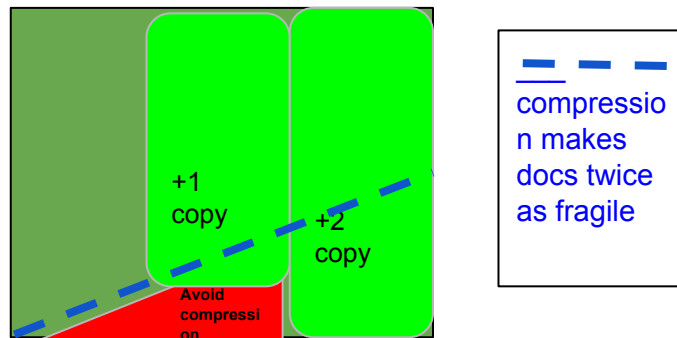


DO compress documents to buy more replications

Compression Shrinks Target & Reduces costs



Compression vs. Repairability: The SWEET Spot



X axis - compressibility; Y is repairability ;
shade by whether reliability is increased; line
plots a fixed proportion reduction of
repairability; overlay line graph of additional
number of copies

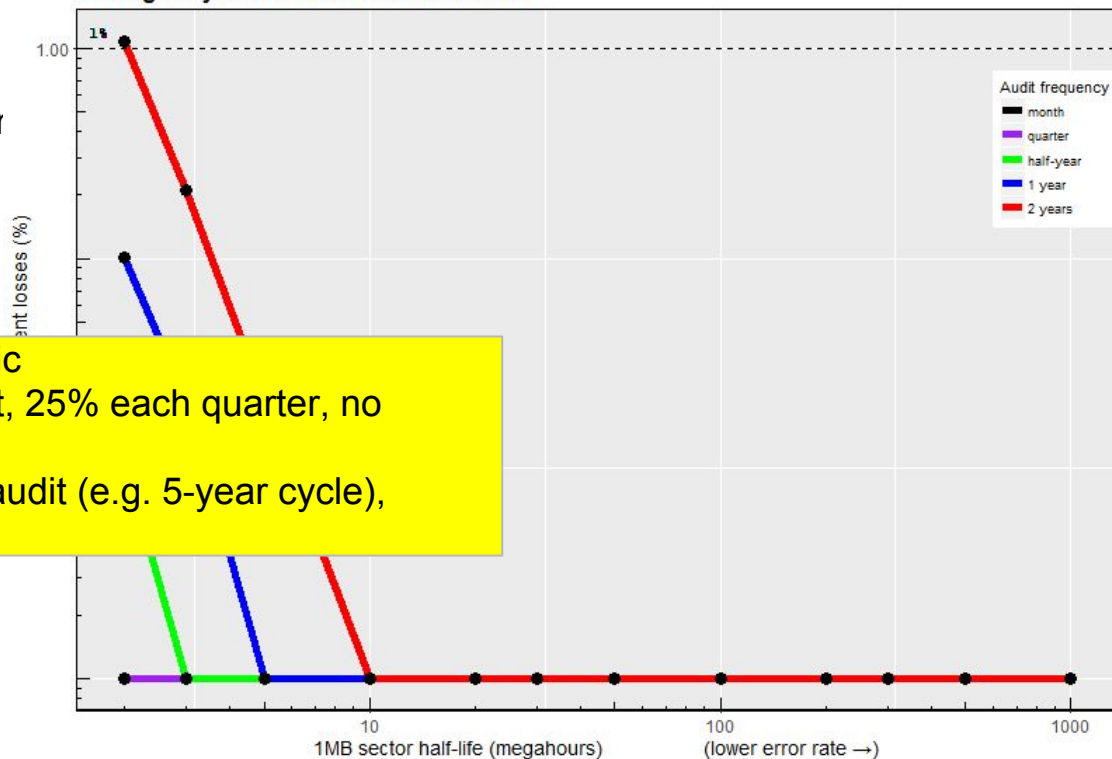
DON'T use Randomized Auditing -- Keep it Systematic

Audit frequency comparisons, copies=5
Increasing auditing frequency beyond annually yields small benefits
but at greatly increased cost in bandwidth

[20% a year without replacement]

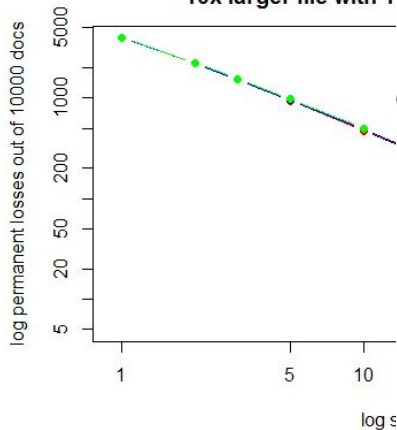


Line 1 annual audit systematic
Line 2 quarterly random audit, 25% each quarter, no replacement over year
Line 3 annual random 20 % audit (e.g. 5-year cycle), replacement every year



DON'T Worry DO be robust

DocSize comparis
10x larger file with 1



Comparisons: larger docs and
corresponding increase in sector
lifetimes yield the same loss rates

Document size	Sector Lifetime (half-life)	Percent of Collection Lost
5	2	15.9
50	20	15.9
500	200	15.9
5000	2000	15.9
5	3	10.9
50	30	10.9
500	300	10.9
5000	3000	11
5	5	6.6
50	50	6.7
500	500	6.7
5000	5000	6.7
5	10	3.3
50	100	3.4
500	1000	3.4
5000	10000	3.4
5	20	1.7
50	200	1.7
500	2000	1.7
5	30	1.1
50	300	1.2
500	3000	1.2
5	50	0.7
50	500	0.7
500	5000	0.7

but document size →

Annotate to show how shifting from 5MB->5000MB
Doc is equivalent to shifting along sector error



Opining

Recommendations

for Memory Institutions

- Use the cloud
- Replicate and verify
- Diversify for server failures
- Compensate for shocks

for Vendors

- Support auditing primitives
- Collect and share loss rates
- Forget 11 nines ...
reveal replication strategy

References

-