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

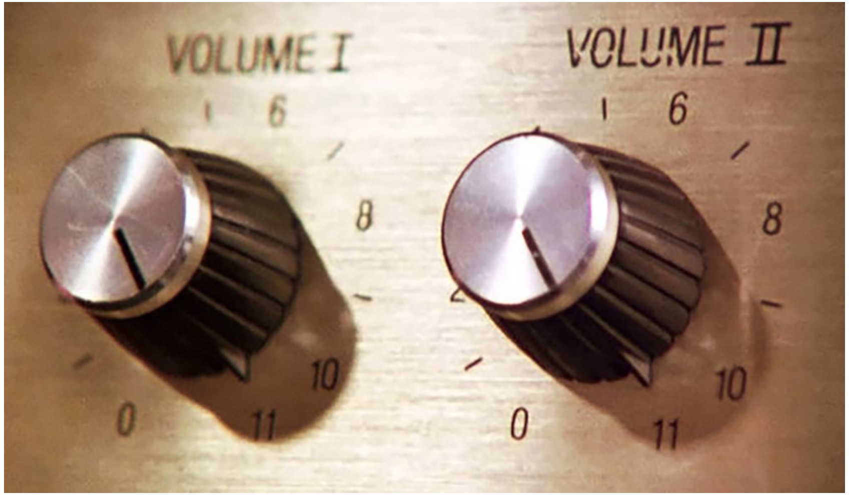
Shifting Economics of Digital Information

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(Loss(C,S*)) | Cost(C,S*)≤ **B**

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

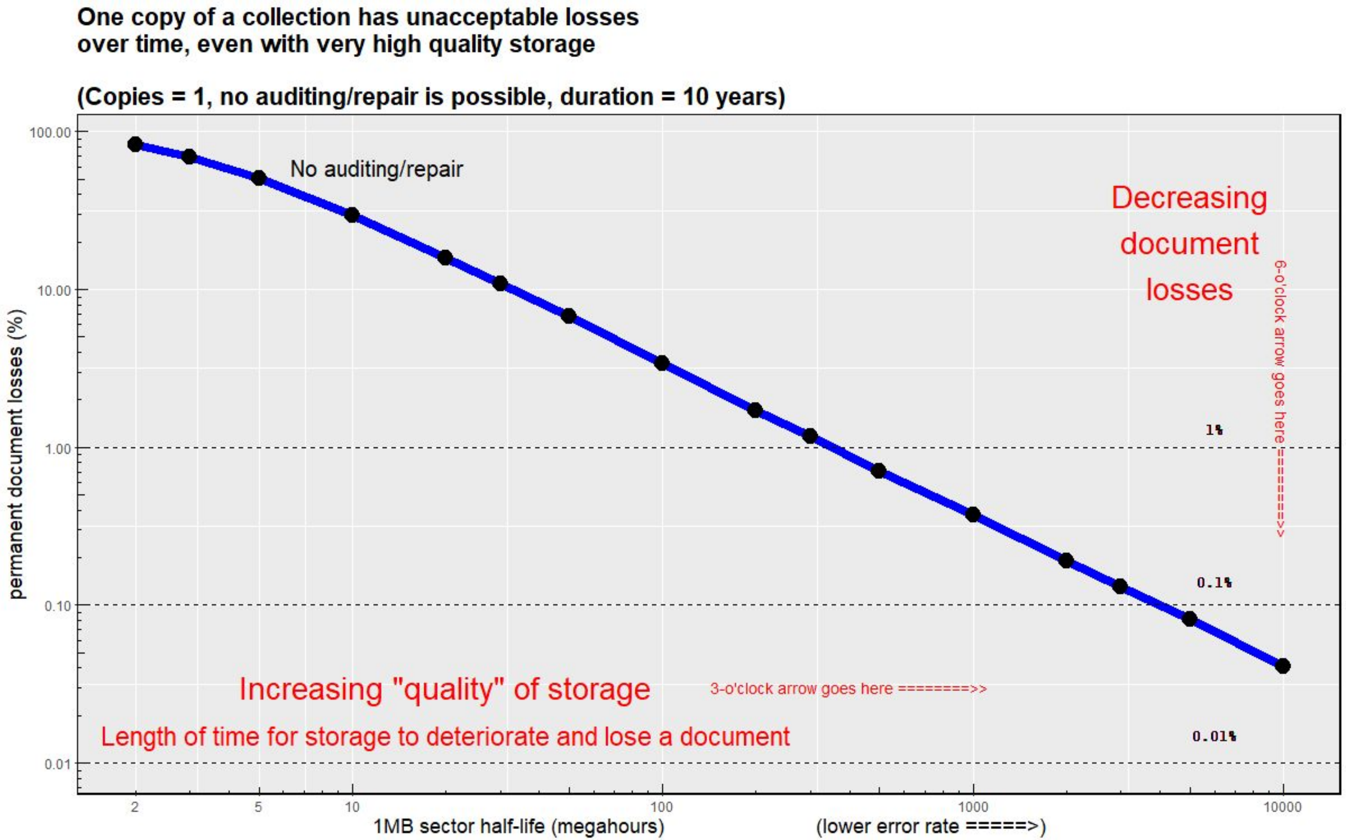
→ **Cost(C,S) =** B1*Replicas + B2*AuditFrequency*Size(C) + B3*Size(C)*Replicas*CompressionFractor(S)

Loss Modeling

Sector	**Corrupts portion of document**	● Detected on audit (silent) ● Exponentially distributed ● Related to storage quality
Glitches	**Environmental Conditions**	● Periodic changes ● Increases sector error rate ● Never directly observable (latent)
Server	**Replica failure**	● Entire replica of collection is lost ● Exponentially distributed
Shock	**Major correlated failure**	● Induces immediate server failure ● May raise rate of server failure

The Big Things

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



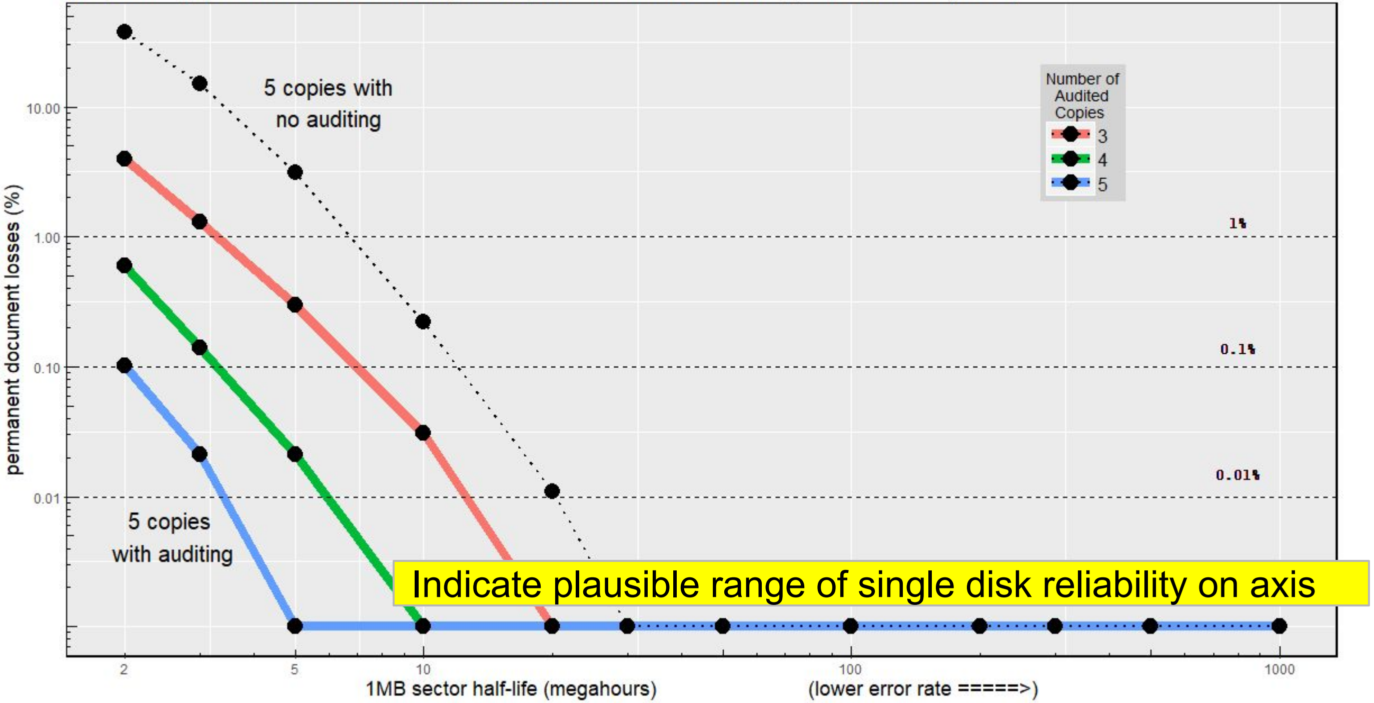
One copy of a collection has unacceptable losses
over time, even with very high quality storage

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

Some Copies + Auditing is better than Many Copies

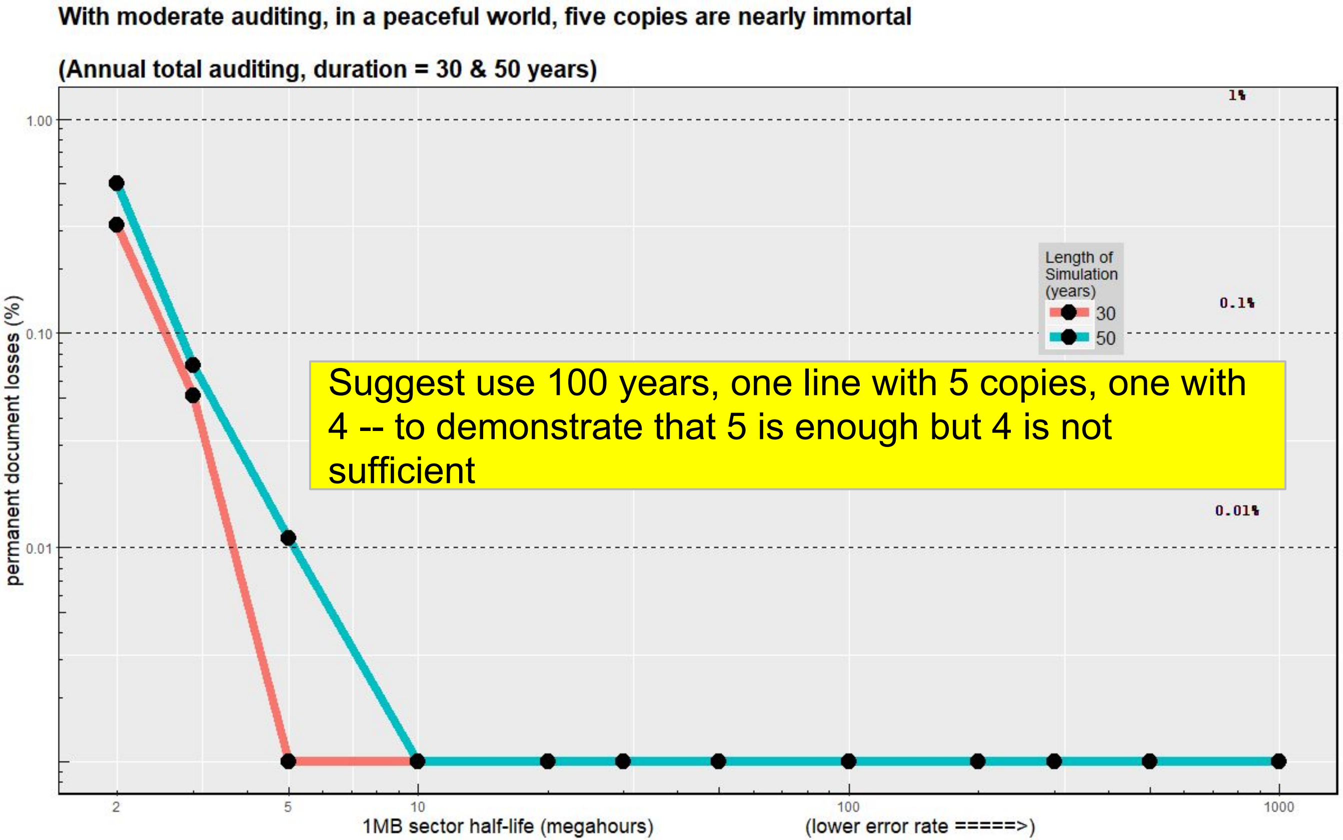
**With regular auditing, only a few copies are required to minimize losses over a wide range.
Failure to audit the collection is worse than keeping
only a small number of audited copies**

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



Indicate plausible range of single disk reliability on axis

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



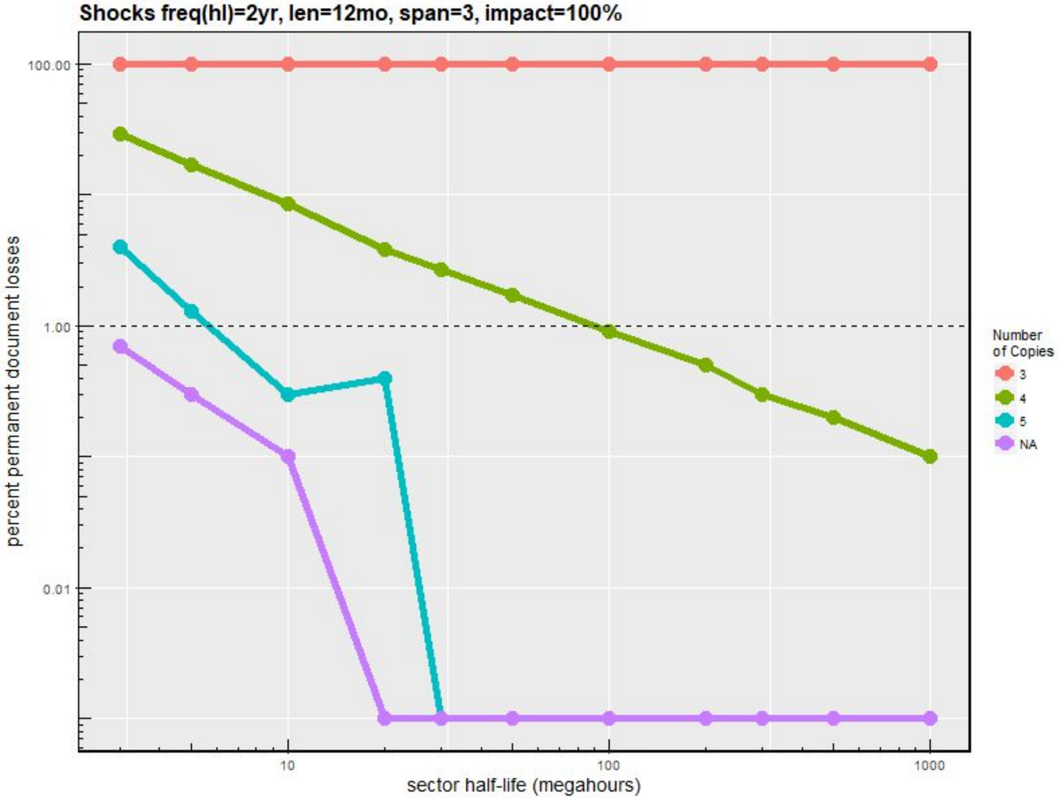
With moderate auditing, in a peaceful world, five copies are nearly immortal

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

Suggest use 100 years, one line with 5 copies, one with 4 -- to demonstrate that 5 is enough but 4 is not sufficient

(With enough copies...)

Sector error doesn't matter, server lifetime does



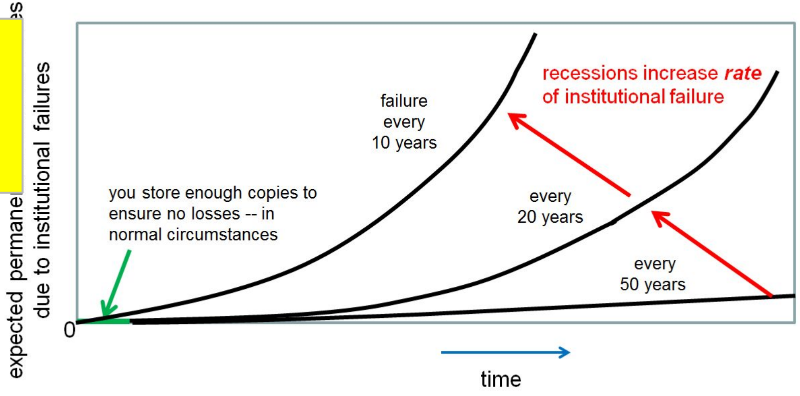
Shocks are Everywhere…

Single server failure?

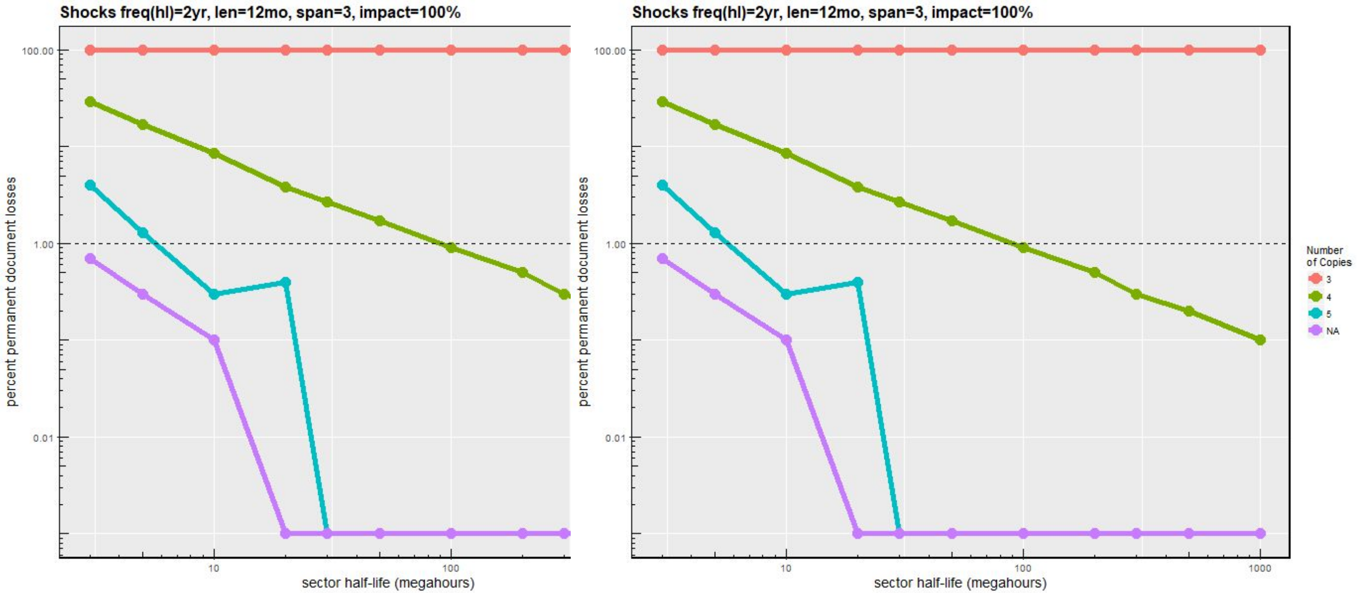
Repression, Encryption Key Loss, Financial

Recession

Companion abstract figure showing sudden corellateoss?



Shocks matter -- even for long-lived servers

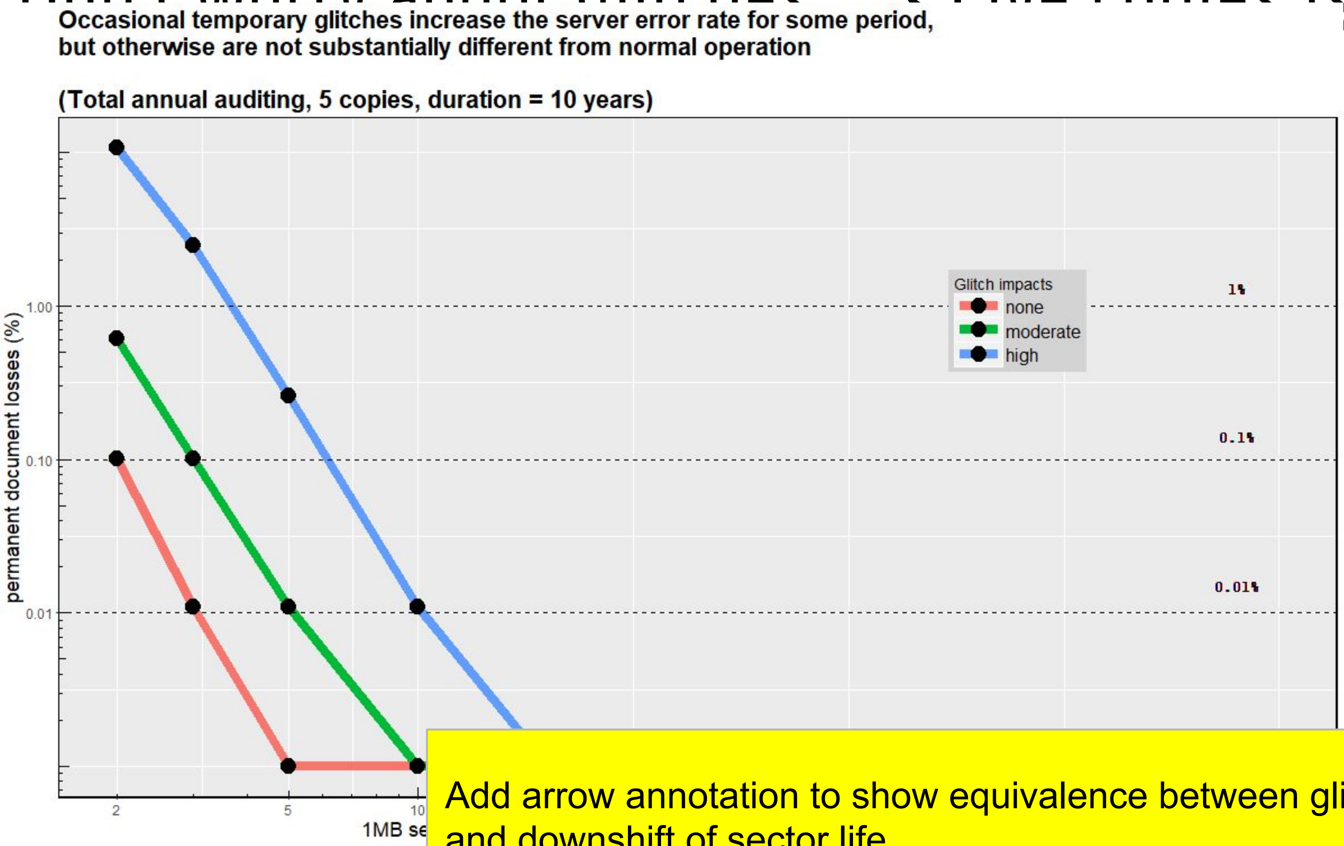


Seven (?) diversified copies will survive a major disaster or minor war

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 → Five copies is enough



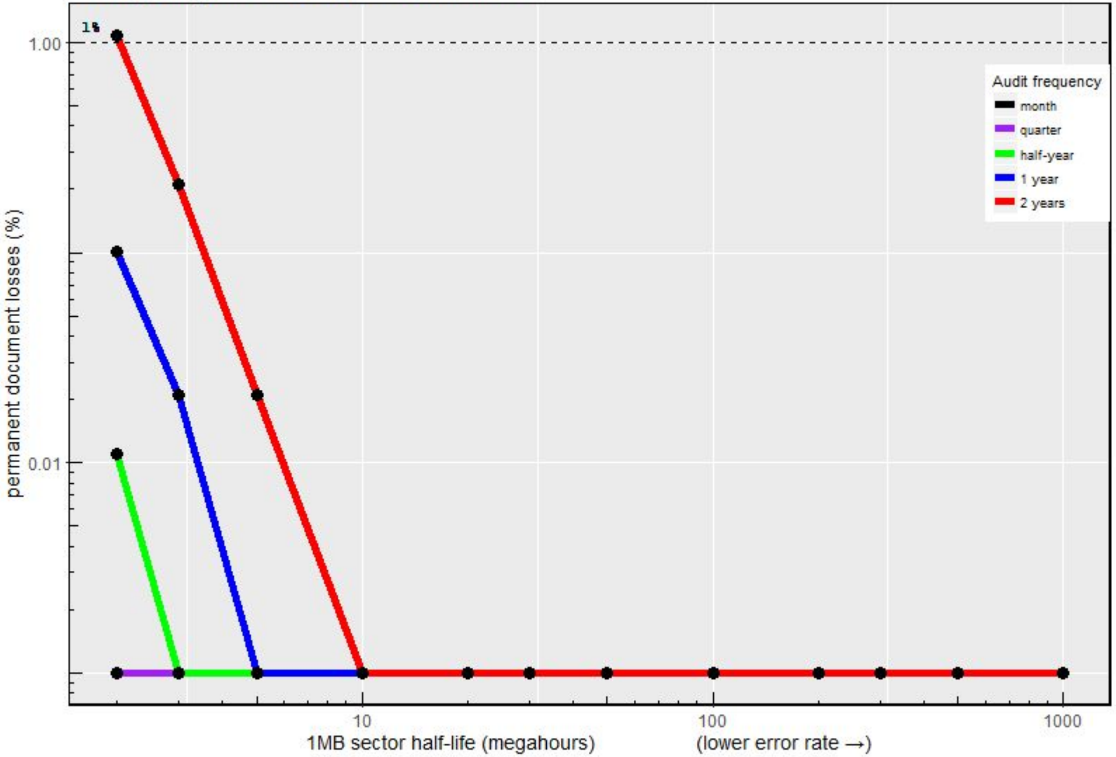
Occasional temporary glitches increase the server error rate for some period, but otherwise are not substantially different from normal operation

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

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

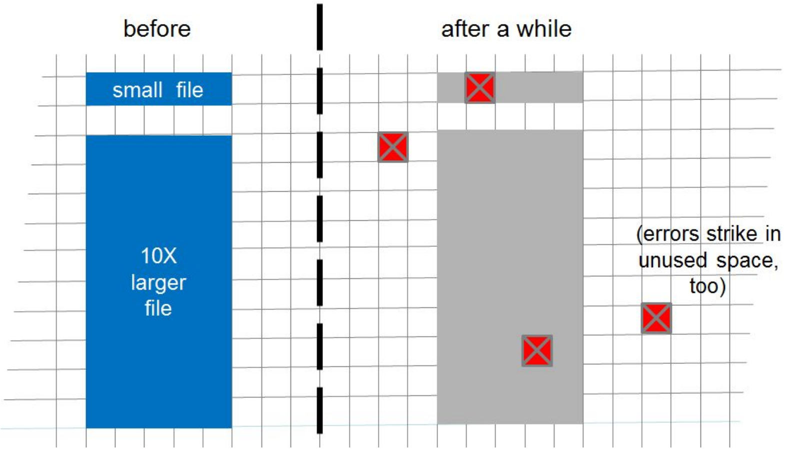
DON’T Worry about auditing frequency -- Annually is Enough

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

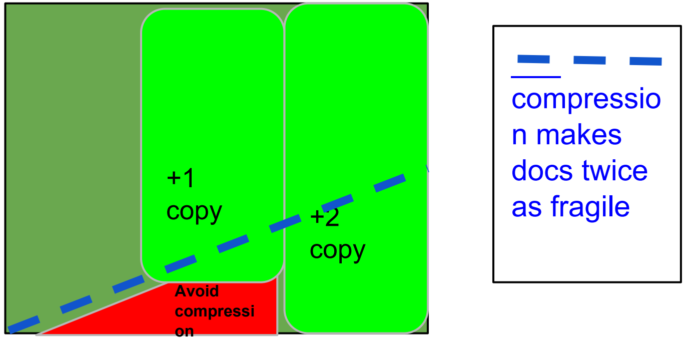


DO compress documents to buy more replications

Compression Shrinks Target & Reduces costs



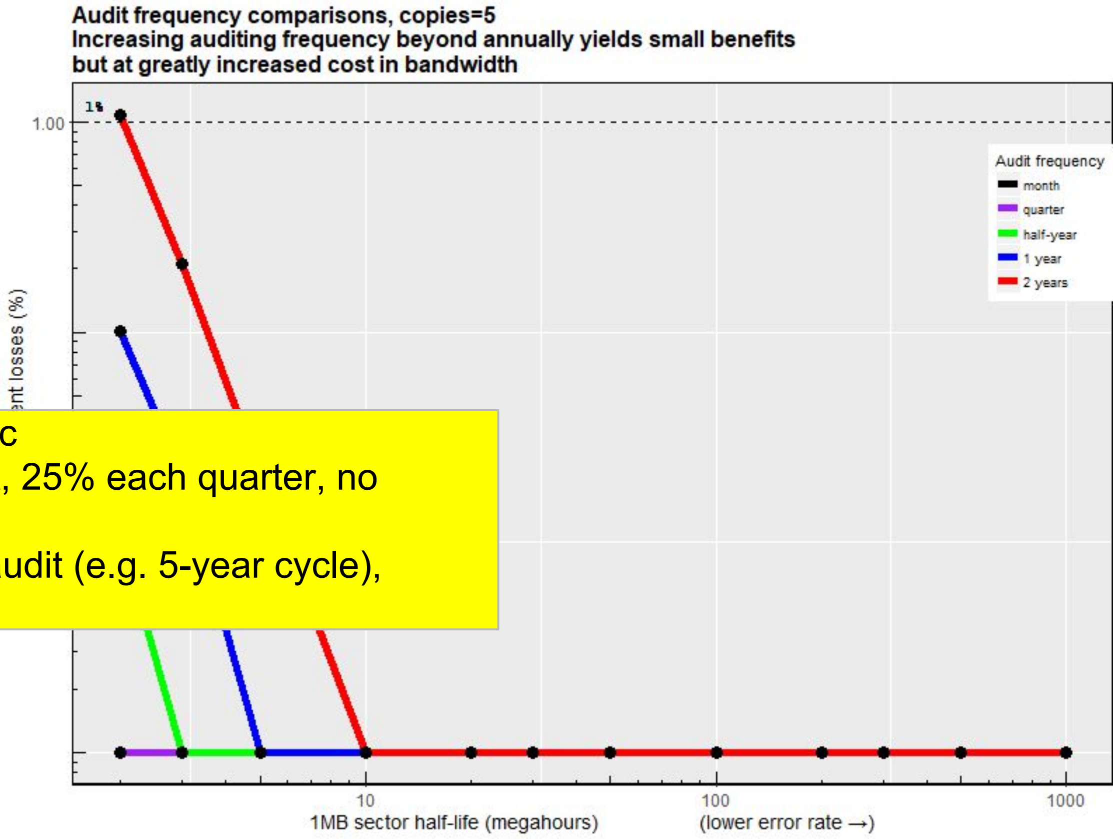
Compression vs. Repairability: The SWEET Spot



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

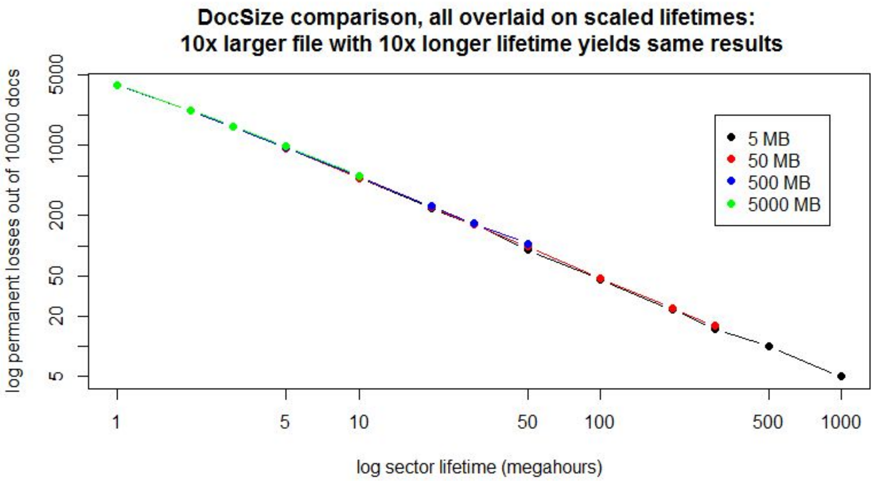
[20% a year without replacemer



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

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 (too much) about document size → DO be robust to sector erros



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

-