Outline for New England NDSA, Friday, 2015-09-25

Problem to Solve

You have a large, valuable, digital document collection

- How many copies do you need to keep it safe?
- On what quality level of servers?
- How often should you audit the servers? Do they still have all the docs?

Not much hard data on which to base policy decisions

We are trying to provide some data, admittedly hypothetical

Assumptions

Everything costs $

- Copies
- Higher quality services to store your docs
    - Data generally not available
- Bandwidth
- Bandwidth for auditing

Trying to provide data you can use for your policies

Basic Data From Which to Extrapolate

Not keyed to any specific problems

Many hints on how to extrapolate from our data to your situations

- Number of docs, doc sizes, storage shelf sizes
- Server failure rates

- Audit strategies

Two Type of Failures

Common: A copy of a document dies on a particular server

Less common: daA server dies, losing all the documents it contains

- "Institutional failure:" fire, flood, war, economic downturn, realignment of purpose, failure of credit arrangements, etc.

All failures are silent (to the client = library)

Auditing is *really* essential

All failures are random, seriously

- They happen at some rate, but are not predictable at all
- Examples of server failure rates, extremely wide spread

Form of the Data

Fixed number of documents

- Scale to your needs

Fixed duration

- Technology changes quickly (weasel words: even ten years is too long)

Number of copies varies, 1 to 10

Reliability of storage servers varies

- Very little real data in this area

Auditing strategies vary

- Frequency, total/partial, random

Document size varies (but doesn't matter)

- The "bigger target" analogy

"Glitches:" from minor A/C failures or bad batch of disks, up to institutional failure

Digital Simulation Programs

Input = error rates, numbers of copies, auditing strategy, etc.

Output = number of documents permanently lost over the life of the test

"Open Source:" will be freely available for others to use, test, verify

"It's just computer time."

Graphs

Ooh, pretty!

- If no auditing, losses by copies and server error rates
- For annual and semi-annual total auditing, same

Preliminary Conclusions

More copies are better (duh!)

Auditing is essential to collection health

- Very frequent auditing is probably overkill

Auditing is expensive (in bandwidth, bytes moved, time)

- We should work toward an efficient cryptographic auditing function

Simple glitches simply increase error rate for a while

Institutional failures are pernicious, but how often do they occur?

- A silent institutional failure reduces the number of redundant copies you have stored
    - Thought you had four copies? Well, for a period of time, you actually had only three.
    - And another failure before the audit would reduce copies to two

- Until you discover the problem (in auditing) and provision a new server

How many copies do you need to limit losses?

- How many to keep likelihood of any permanent loss under some percentage?

    - 5 per cent, 1 per cent, 0.1 per cent?

- For institutional failures, particularly correlated, how many copies to keep likelihood of total loss under some percentage?

[Ten-ish slides is probably too many for a short session. How long is the session? How active is the audience likely to be in questioning? Relegate some of the slides to "backup." You know the audience; you pick.]