Protecting Digital Content Through Failure Analysis and Modeling

Library of Congress Meeting on Storage Architectures

Jeff Robinson, Ph.D.

AQI | Accelerated Quality Improvement

Introductions

- Dr. Jeffrey Robinson, VP Technology, Accelerated Quality Improvement; Principal Consultant and Investigator
- Scott Rife, Library of Congress: OCIO

Initial Objectives

- In the fall of 2014, NAVCC initiated a project to assess digital integrity of its archives using failure analysis and modeling
  - "Develop a mathematical model that provides a likelihood percentage and confidence intervals given our current infrastructure and … from this model … to be able to evaluate the relative merits of improving our infrastructure"
  - That is, create a descriptive/predictive model of availability/reliability that allows sensitivity analysis and 'what-if' analysis

Overview of Projects

- Interview Stakeholders/Team
- Conduct initial FMEA – Failure Mode Effects Analysis
- Develop initial list of failure modes of greatest concern
- Develop taxonomy of failures and results/effects
- Research Failure Rates and Reliability statistics
- Build Initial Model
- Evaluate model; refine
- Conduct Process FMEA
- Develop detection model
- Prepare findings and recommendations

Some initial concerns

- Silent Corruption
- Corruption of data on HDD
- Loss or corruption of data on tape media

- Effectiveness of different strategies on reducing probability of data loss events

Why Simulation

- Takes too long to determine experimentally
- Too much risk to determine empirically
- Need a model that can predict results
- Need a model that can combine influence of multiple factors
- Need What-if analysis
- Some envisioned solutions/alternatives are too expensive to test
- Desire a model that others can also use

Taxonomy of Failure types

- Type 0 – Content inaccessible but without data loss
- Type 0 – Content accessible but slow
- Type 0 - Online content not available go to primary tape
- Type 1 – Primary data loss; backup media available
- Type 2 – Primary and Backup loss; need to re-digitize source
- Type 3 – All local copies lost; original source no longer available (possible permanent)
- Type 4 – No copies of source available worldwide; permanent
- Type 5 - Multiple files/tapes lost

Type of Failure Modes

- Procedural
  - Metadata Error -setup
  - Procedural Deletion HDD
  - Procedural Deletion
  - Lost Tape
- MAVIS Errors (Collections Management DB for MBRS)
  - Corrupt Index
  - Manual Deletion MAVIS
  - HD Failure
- Tape Drives And HW
  - Read Error On Transfer
  - Tape Media – Repositor
  - Bad Media
  - Broken Tape
  - Bad Write Before Sha1
  - Bad Write During Copy1
  - Bad Drive
- Network
  - Network Storm
  - DDOS
- Miscellaneous
  - Malware
  - Write Error At Outage
  - Disk Error Write(Failure)
  - RAID Error
  - Hard Drive
- Human Error
- Other
  - Malicious acts
  - Hack/Ransomware
  - Virus
  - Fire/Flood
  - Acts of God (earthquake, hurricane, etc.)

FMEA

LOC NAVCC FMEA						
Item	Component	Potential Failure Mode	Potential Effects of Failure (Symptoms)	Potential Cause of Failure	Failure Type (0-5)	RPN
80	procedural	erroneous data retrieved	erroneous data retrieved	mavis vs file system	Type 4 – no copies of source available worldwide;	288
77	procedural	erroneous data retrieved	erroneous data retrieved	metadata file to tape	Type 2 – Primary and Backup loss; need to	216
78	procedural	erroneous data retrieved	erroneous data retrieved	metadata context	Type 2 – Primary and Backup loss; need to	216
4	procedural	erroneous data retrieved	erroneous data retrieved	wrong metadata	Type 2 – Primary and Backup loss; need to	168
60	procedural	erroneous data retrieved	erroneous data retrieved	mismatch content data	Type 2 – Primary and Backup loss; need to	144
62	procedural	erroneous data retrieved	file not found	referential integrity	Type 2 – Primary and Backup loss; need to	144
42	environmental	system down	delays?	problems with restarts	Type 0 - online not available go to primary tape	144
44	tape	unable to retrieve file	unable to read tape	tape degradation	Type 1 – Primary data loss; backup available	140
70	external	various	various	disgruntled employee	Type 4 – no copies of source available worldwide;	126
15	mechanical	unable to retrieve file	tape not found/damage tape	bad robot	Type 1 – Primary data loss; backup available	120
3	procedural	unable to retrieve file	tape not found	lost tape	Type 1 – Primary data loss; backup available	120
30	environmental	CPU down	unable to access system	external power	Type 0 - online not available go to primary tape	120
41	environmental	CPU down	unable to access system	no UPS	Type 0 - online not available go to primary tape	120
58	procedural	unable to retrieve file	file not found	corrupted index	Type 0 - online not available go to primary tape	120
59	procedural	erroneous data retrieved	file not found	bad pointers	Type 0 - online not available go to primary tape	120
75	Virus		various	virus to ORACLE or OS	Type 2 – Primary and Backup loss; need to	112
6	procedural	files deleted	erroneous data retrieved	accidental deletions	Type 2 – Primary and Backup loss; need to	108
10	mechanical	unable to retrieve file	tape not found	bad label	Type 1 – Primary data loss; backup available	105
79	procedural	unable to retrieve file	tape not found	stolen tape	Type 1 – Primary data loss; backup available	105
71	external	system down	various	suicide bomber	Type 4 – no copies of source available worldwide;	96
76	Virus	system down	various	ransomware	Type 3 – Copies lost; original source no longer	96
34	environmental	Facility damage/power	unable to access system	fire	Type 2 – Primary and Backup loss; need to	96
43	environmental	system down	unable to access system	network failure	Type 0 - online not available go to primary tape	96
63	system	system down	unable to access system	network down	Type 0 - online not available go to primary tape	96
64	system	CPU down	unable to access system	power outages (cable short)	Type 0 - online not available go to primary tape	96
72	external	system down	unable to access system	EMP	Type 2 – Primary and Backup loss; need to	84
31	environmental	Electrical damage	unable to access system	lightning	Type 0 - online not available go to primary tape	84
45	tape	unable to retrieve file	unable to read tape	bad media	Type 0 - accessible but slow	84
5	procedural	erroneous data retrieved	erroneous data retrieved	wrong title	Type 0 - online not available go to primary tape	72
8	procedural	files deleted	various	rm *	Type 0 - online not available go to primary tape	72
11	mechanical	unable to retrieve file	unable to read tape	tape windup	Type 0 - online not available go to primary tape	72
14	mechanical	unable to retrieve file	unable to read tape	broken tape	Type 1 – Primary data loss; backup available	105
40	environmental	CPU down	unable to access system	overtemp (cooling failure)	Type 0 - online not available go to primary tape	64
69	supply chain	unable to retrieve file	delays	people resource	Type 0 - online not available go to primary tape	64
61	procedural	erroneous data retrieved		cannot delete files in mavis	Type 0 - online not available go to primary tape	64

Reliability Data

- Sources
  - Industry data
  - Vendor data
  - Historical data
  - Anecdotal data
  - Other data sources
- Reliability Data Issues
  - Incomplete data
  - Vendor Bias
  - Units
    - MTBF

Reliability Data - Units

- Probabilities (for time periods, day, weeks, months, years)
- Probabilities by IO (KB, MB, GB, or even MiB)
- MTBF or Mean time for Service
- Occurrences per opportunity (DPMO or Six Sigma levels)
- Percent of failures of service lifetime
- Probabilities per specific events

- All probabilities were standardized to probability per day (based on typical ingestion and usage at NAVCC)

Simulation Tool

Failure Model 4

Background	Assumptions		
Read to write ratio - 1w 3r	digitize media, sha, tape 1, tape 2, hdd, read from there		
1tb tape migrate to 5tb tape			
another migration in 5 years			
total size	5.5PB	per repository	5,500,000,000,000,000
size of tape	5000000000000	1TB	
number of tapes	5500	per repository	
number of files	1,200,000	per repository	bytes
average files per tape	218.18		
average size per file	4,583,333,333	about 5GB	
size of online store	100TB	want to extend to 200TB	
adds per year	1-2PB/year		
adds per day	10,000,000,000,000	10TB/day	assume 200days/year
new files per day	2,182		
tapes add per day	2	5-10/day	
total I/O per day	40,000,000,000,000	40TB/day	x4+users
number of drives per repository	5		row 37
files or tapes accessed / day	28	5000files/year	plus2 new tapes per over 200 days
usage rate - retrieval & ingestio	125,000,000,000	5gb/file*5000file	0.002739726 1000-5000 files /year

Instructions on use of Failure Model

- The model only works with active Crystal Ball License
- Green cells are assumptions (probability distributions used by Monte Carlo Analysis application)
- Turquoise fields are outcomes (forecasts generated by the model)
- The pink cell will stop the simulation when a failure (value = 1) appears in the specified row of column F
  Otherwise the simulation will continue until the specified number of trials is completed
- Columns F and G are values of current trial
- Column H is the sum of the failure for all iterations
- Column I shows sums of type 1, 2, and 3 failures for the current trial

Failure types

Type 0 – inaccessible but without data loss
Type 0 - accessible but slow
Type 0 - online not available go to primary tape
Type 1 – Primary data loss; backup available
Type 2 – Primary and Backup loss; need to re-digitize source
Type 3 – Copies lost; original source no longer available (possible permanent)
Type 4 – no copies of source available worldwide; permanent
Type 5 - multiple files/tapes lost

Stop simulation when failure(1) appears in

Cell column F and Row = 1 or 1

0

Class of failure	Failure mode	Probability	Probability per day	Type	v1	v2	Number Failures	
Procedural	metadata error	1/million files	0.00000000274	2	0	0	0	0
	procedural deletion HDD	10%of HD failure	0.00000329	1	0	0	0	
	procedural deletion tape	1/10yrs	0.000273973	1	0	0	24	0
	lost tape	1/5yr	0.000547945	1	0	0	51	
								0
Mavis error	corrupt index	1/million files	0.00000000	1	0	0	0	
	manual deletion mavis	1/4 years	0.000684932	1	0	0	65	
	hd failure	6% of 5 year life	0.00003288	1	0	0	3	
Tape Drives and HW	read error on transfer	10-19	0.0000010000	2	0	0	0	
Tape media - repository1	bad media1	1/2500 tapes	0.0008	1	0	0	82	
	broken tape1	1/20ktapes	0.000005	1	0	0	0	
	bad write before sha1	10-16	2.18182E-13	2	0	0	0	
	bad write during copy1	10-16	2.18182E-13	2	0	0	0	
	misaligned heads1	.01%/year	2.73973E-06	1	0	0	0	
	bad drive	0.0000329		1	0	0	1	
Tape media - repository 2	bad media2	1/2500 tapes	0.0008	1	0	0	68	
	broken tape2	1/20ktapes	0.000005	1	0	0	1	
	bad write before sha2	10-16	2.18182E-13	2	0	0	0	

Outcomes/Results Number of Errors by Type		
Iterations (or	= 273.97	100000 years)
Type 0	0.03069	3069
Type 1	0.00311	311
Type 2	0.00002	2
All Types		3382

Total bad tapes 152

If two bad tapes experience Type 1 failures
What are the odds that they are on the same tape in the both repositories

Actual percentage of commonality is given by the expression:
=1-(FACT(11000)/(FACT(11000-L43))/(11000^L43))
However, 107 is the maximum value excel and use in a factorial function

# bad tapes	chance they are the same
2	0.04%

Rubric Conversion

Iterations	Years
1,000,000	2,730
100,000	273.97
36500	100
18250	50
10,000	27.3
7300	20
3650	10
1825	5
365	1
8760	hrs per year

Excel / Crystal Ball
Monte Carlo Simulation)

Findings and Conclusions

- Potential data loss events due to HW and SW failure are less than suspected
- Potential data loss events due to human error, human action or procedural error is greater than was suspected
- Current fault tolerant strategies are extremely effective
  - Hash Digest
  - Multiple copies in separate repositories

Some Discoveries

- Detection is critical to managing risks and forestalling serious data loss events
- Detecting errors as soon as possible after they occur is key o minimizing errors that could become serious data loss events

- Need to develop new metrics and monitors that can be used as precursors/predictors of data loss events (e.g. disk errors, procedural errors)

Systems Control Theory

- Dynamic Systems are unstable
- Good systems go bad; Bad systems get worse
- If left uncorrected, errors accumulate and grow over time (entropy wins)

- What keeps dynamic systems stable is "feedback" – The detection and correction of errors

Next steps

- Additional Areas of investigation (opportunities for further research)
- Develop and maintain a *failure database*
- Share and refine model – and findings with other similar facilities
- Extend model using real (historical) data
- Share and standardize metrics with other facilities

Summary

- Protecting digital content
  - Failure analysis
  - Quantitative Risk Analysis
  - Simulation and Modeling

Q & A