STANFORD UNIVERSITY LIBRARIES

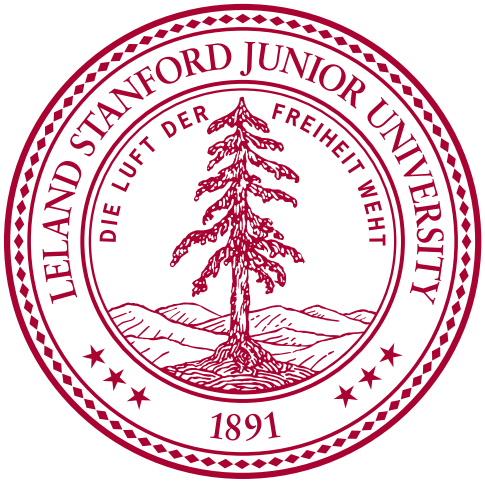
Digital Library Systems and Services (DLSS)

Report of Monte Carlo Simulation of Digital Object Placement on Tape

Version 1.0 Friday, April 26, 2013

STANFORD

UNIVERSITY



Author: James Simon (jjsimon@stanford.edu)

Overview

This document discusses analysis of various aspects of making copies of digital objects to tape. A model was created that included the number of copies, number of tapes, number of versions, and scattering across tapes.

The Monte Carlo method was used for the analysis and will be described in detail later. Basically, a Monte Carlo simulation models what happens in the real world and applies probability to particular events. The simulation runs (thousands of times) in order to perform a deterministic computation on the events and aggregate the results.

Since we are interested in determining how reliably we can recover digital objects from tape, the model was designed to answer these questions:

1. What impact does having more copies of a digital object have?
2. Since SDR uses forward versioning, what impact does having more versions on tape have?
3. Is it better to have versions of a digital object all on the same tape, or scattered across multiple tapes?

Monte Carlo Method applied to replication on tape

As mentioned earlier, the goal of a Monte Carlo simulation (henceforth just simulation) is to model the real world.

Reading a Digital Object

So the simulation in our case is to determine if a digital object version can be read using a particular replication strategy. In order to read the last version of a digital object, we must read all previous versions and the following applies:

1. Reading tapes – an attempt to read is successful given a probability that a tape is good or bad
    a. Once a tape is bad, it is bad forever, i.e. it cannot be used to read other versions on that tape.
    b. All tapes are good in the beginning, but could turn bad when read.
2. Reading a version – when attempting to read a version (say version 3) of a digital object, a probability for success is applied to the read, this is the probability that the version is corrupt.
3. If a version is corrupt on a good tape, we try again on the next tape, if there is a next tape.

4. If any of the versions in a versioned digital object are bad, this is considered an unsuccessful attempt. If all are read, then it is a successful attempt.

The simulation keeps track of successful digital object retrievals and failed tapes.

Tape Replication Strategies

Earlier we asked these three questions:

1. What impact does having more copies of a digital object have?
2. Since SDR uses forward versioning, what impact does having more versions on tape have?
3. Is it better to have versions of a digital object all on the same tape, or scattered across multiple tapes?

To answer question 1, we have to simulate number versions, number of tapes and number of copies. So for example, we might have a digital object that has 4 versions, and two copies. This requires a minimum of 2 tapes. This would be an example of a single replication strategy. Namely:

- Two copies on tape
- All copies aggregated

Parameterized it might look like this:

version_num = 4
tape_num = 2
copy_num =2

Versions: ["v1", "v2", "v3", "v4"]

Tape 0:
element 0 = v2
element 1 = v4
element 2 = v1
element 3 = v3
Tape 1:
element 0 = v1
element 1 = v3
element 2 = v2
element 3 = v4

Note that the order written to tape is random.

If we change this basic strategy to include 3 or more tapes, it would change the layout on tape and affect the model.

For example:

Versions: ["v1", "v2", "v3", "v4"]

Tape 0:
 element 0 = v1
 element 1 = v3
 element 2 = v2
 element 3 = v4
Tape 1:
 element 0 = v2
 element 1 = v1
Tape 2:
 element 0 = v4
 element 1 = v3

This is how we model question 3, i.e. versions disaggregated and scattered across tapes. Note that the placement for all digital objects is random, but not duplicated on the same tape. This models how we would do this in the real world.

So, finally if we want to answer question 2, we could sent the number of versions to a high number and keep the other parameters the same as previous simulations.

For example 10 versions, 3 tapes and 2 copies might look like this:

Versions: ["v1", "v2", "v3", "v4", "v5", "v6", "v7", "v8", "v9", "v10"]
Tape 0:
element 0 = v1
element 1 = v3
element 2 = v6
element 3 = v7
element 4 = v9
element 5 = v10
Tape 1:
element 0 = v4
element 1 = v5
element 2 = v8
element 3 = v2
Tape 2:
element 0 = v2
element 1 = v6
element 2 = v7
element 3 = v9
element 4 = v10
element 5 = v1
element 6 = v3
element 7 = v4
element 8 = v5
element 9 = v8

Methodology of Simulation runs

For each strategy the simulation runs 250,000 times, with 5 iteration steps, with randomly generated placement and distribution across tapes (if number of tapes is grater than the number of copies.

With regards to the probabilities for reading tapes and versions, they are set arbitrarily high, e.g. 1% chance for failure. This is not to skew results for any particular model, but to introduce relative failure rates for comparison purposes. Thus, these simulations are not for failure prediction, but for comparison of replication strategies.

Following this general description of the simulation and methodology are the results of various runs using different strategies, along with analysis.

Simulation Results

Impact of version number on reliability

The following simulations were run.

version_num = 10
tape_num = 2
copy_num =2

50,000 x 5 iterations

successful reads = 47745
Number of bad tapes = 0
successful reads = 47865
Number of bad tapes = 2
successful reads = 47777
Number of bad tapes = 0
successful reads = 47768
Number of bad tapes = 0
successful reads = 47793
Number of bad tapes = 0

After 5 iterations, average is: 47789
standard deviation
40.77057762651886
Percentage success: 95.578%

version_num = 1
tape_num = 2
copy_num =2

50,000 x 5 iterations

successful reads = 49940
Number of bad tapes = 0
successful reads = 49924
Number of bad tapes = 0
successful reads = 49925
Number of bad tapes = 0
successful reads = 49922
Number of bad tapes = 0
successful reads = 49932
Number of bad tapes = 0

After 5 iterations, average is: 49928
standard deviation
6.6211781428987395
Percentage success: 99.856%

Version = 1
Tape_num = 1
Copy_num = 1

successful reads = 48024
Number of bad tapes = 0
successful reads = 47979
Number of bad tapes = 0
successful reads = 48107
Number of bad tapes = 0
successful reads = 48052
Number of bad tapes = 0
successful reads = 47978
Number of bad tapes = 0

After 5 iterations, average is: 48028
standard deviation 48.44378185071847

Percentage success: 96.056%

Version = 10
Tape_num = 1
Copy_num =1

successful reads = 33297
Number of bad tapes = 0
successful reads = 33283
Number of bad tapes = 2
successful reads = 33345
Number of bad tapes = 0
successful reads = 33283
Number of bad tapes = 0
successful reads = 33352
Number of bad tapes = 0

After 5 iterations, average is: 33312
standard deviation 30.31831129861952

Percentage success: 66.624%

In every case, increasing the number of versions has an impact on reliability of successfully reading a version. The simulation modeled after a Forward versioning scheme. Note also that increasing number of copies also seems to improve reliability.

Impact of number of copies on reliability

version_num = 2
tape_num = 3
copy_num =3
successful reads = 49990
Number of bad tapes = 0
successful reads = 49994
Number of bad tapes = 0
successful reads = 49995
Number of bad tapes = 0
successful reads = 49988
Number of bad tapes = 0
successful reads = 49994
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 49992
standard deviation
2.7129319932501073
Percentage 99.9844

version_num = 2
tape_num = 3
copy_num =2
successful reads = 49815
Number of bad tapes = 0
successful reads = 49836
Number of bad tapes = 0
successful reads = 49815
Number of bad tapes = 0
successful reads = 49820
Number of bad tapes = 0
successful reads = 49809
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 49819
standard deviation 9.186947262284681
Percentage 99.638

version_num = 2
tape_num = 3
copy_num =1

successful reads = 46153
Number of bad tapes = 0
successful reads = 46021
Number of bad tapes = 0
successful reads = 46217
Number of bad tapes = 0
successful reads = 46133
Number of bad tapes = 2
successful reads = 46057
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 46116
standard deviation 69.80658994679513
Percentage 92.2324

version_num = 5
tape_num = 3
copy_num =3

successful reads = 49951
Number of bad tapes = 0
successful reads = 49956
Number of bad tapes = 0
successful reads = 49957
Number of bad tapes = 0
successful reads = 49958
Number of bad tapes = 0
successful reads = 49961
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 49956
standard deviation
3.2619012860600183
Percentage 99.9132

The number of copies improves reliability of a retrieval, increasing the number of versions has a slight impact, but more copies overwhelm version number impact.

Impact of number of tapes on reliability

version_num = 10
tape_num = 10
copy_num =2

successful reads = 48950
Number of bad tapes = 0
successful reads = 48840
Number of bad tapes = 0
successful reads = 48937
Number of bad tapes = 0
successful reads = 48836
Number of bad tapes = 0
successful reads = 48810
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 48874
standard deviation
57.33968957013981
Percentage 97.74919999999999

version_num = 10
tape_num = 5
copy_num =2

successful reads = 48517
Number of bad tapes = 0
successful reads = 48509
Number of bad tapes = 0
successful reads = 48551
Number of bad tapes = 0
successful reads = 48467
Number of bad tapes = 0
successful reads = 48464
Number of bad tapes = 0
Iterations = 50000
After 5 iterations, average is: 48501
standard deviation
32.69005965121508
Percentage 97.0032

version_num = 10
tape_num = 2
copy_num =2

successful reads = 47806
Number of bad tapes = 0
successful reads = 47810
Number of bad tapes = 0
successful reads = 47786
Number of bad tapes = 0
successful reads = 47806
Number of bad tapes = 0
successful reads = 47867
Number of bad tapes = 2
Iterations = 50000
After 5 iterations, average is: 47815
standard deviation
27.320322106446696
Percentage 95.63000000000001

The impact of number of tapes shows that increasing the number of tapes increases reliability. There is a correlation to the number of versions/copies as the copies/versions must be spread across the tapes. This will not be a problem for SDR as the number of items will fill tapes.

Conclusion

More copies are much better, for forward versioning, there is a penalty as the number of versions of a digital object increase, and finally, spreading versions across more tapes can marginally improve reliability.