Reliability modelling for long term digital preservation

Panos Constantopoulos[1,2], Martin Doerr[2], Meropi Petraki[2]

[1]Athens University of Economics and Business
[2]Institute of Computer Science, FORTH
{panos,martin,petraki}@ics.forth.gr

Abstract.

Digital material is vulnerable to loss and corruption as it is stored in magnetic and optical media that can fail because of exposure to heat, humidity, airborne contaminants or because of faulty reading and writing devices. Organizations whose mission encompasses the preservation of information and cultural heritage have very high reliability requirements for their digital preservation systems. The problem essentially is to achieve given reliability goals with minimal cost. In an on-going project we study the effectiveness of various system configurations for digital preservation by applying analytical modelling techniques to evaluate the reliability of those configurations. The results are expected to be useful in supporting such decisions as how many copies to retain for each document, how frequently to check these copies for corruptions, what storage schemes are most effective for data preservation, what strategies are most effective for error recovery and repair, how many sites should be used in order to prevent loss from natural disasters such as flood, fire, earthquake. Our work includes sensitivity analysis of a set of reliability criteria with respect to model parameters in representative cases. This can prove especially helpful in identifying dominant factors with respect to the reliability and cost of each configuration. Preliminary results exhibit a certain degree of counter-intuitiveness, which adds value to the model. In this paper we present the main points of the model and outline some policies suggested by the analysis so far.

Introduction

In this paper we analyze realistic back-up strategies to preserve digital contents over the very long times for which museums and libraries would like to keep our documented social memory.

Motivation

Digital material is vulnerable to loss and corruption as it is stored in magnetic and optical media that can fail because of exposure to heat, humidity, airborne contaminants or because of faulty reading and writing on devices. Organizations whose mission encompasses the preservation of information and cultural heritage

have very high reliability requirements for their digital preservation systems. A larger museum or library holding between a 100.000 to several million objects may not want to loose more than a dozen documents per year. This number equals to the astonishing requirement not to loose more than about 1% of its holdings in a thousand years. Scholars still mourn the loss of the library in Alexandria two thousand years ago. During such immense time spans, most of the data carriers have been replaced many times. Hardly any piece of literature from antiquity is preserved on the initial carrier.

The problem essentially is to achieve given reliability goals with minimal cost. It does not matter, that current technology will also have been replaced many times. The only assumption is that there are data carriers used still then. Let's assume we apply an optimal strategy with a technology that holds for ten years from now. In these ten years, we have met the reliability goal. Then we apply the same method of analysis presented here to the next generation technology and apply the results for the next period of technology. So we can infinitely proceed with the same desired reliability factor independent of the technology.

Current work in digital preservation has concentrated on the readability of formats in the future, and discusses data replication and migration as if it would be an alternative. Over the relevant time-spans, it is a simple necessity. The longest human memory known to the authors of this paper is the oral tradition of the Qksan tribes in British Columbia that is reported to go back more than 10.000 years, by highly disciplined information control, replication and migration among human brains [13].

We are therefore studying the effectiveness of various system configurations for digital preservation by applying analytical modeling techniques to evaluate the reliability of those configurations. The applied method is useful in supporting such decisions as how many copies to retain for each document, how frequently to check these copies for corruptions, what storage schemes are most effective for data preservation, what strategies are most effective for error recovery and repair, how many sites should be used in order to prevent loss from natural disasters such as flood, fire, earthquake. We have conducted sensitivity analysis of a set of reliability criteria with respect to model parameters in representative cases. Sensitivity analysis has proved especially helpful in finding 'major' factors that influence the reliability and cost of each configuration. In this paper we present first results. The examples show that intuition may fail to estimate such factors. We discuss the general relevance of these results for data replication strategies.

Related Work

Following [6], the major threats to digital information are:

- Media decay and failure
- Access Component Obsolescence
- Human and Software Errors
- External events

There is vast literature about how to address these risks. There are preventive measures and recovery strategies.

The methods discussed to address Access Component Obsolescence are either migration [20,15,7,10], preservation of obsolete technology, preservation of know-

how together with emulation of obsolete technology [16] and finally encoding standards that are easy to interpret. Still the core of our cultural and scientific memory is in the form of text and data elements that can be interpreted by humans if the schema is known. For this class of information, XML seems to show a way out of S/W obsolescence. For obsolescence of the data carrier technology, migration seems to be the only long-term alternative.

In order to reduce media decay and system failure, control of environmental conditions is studied. [21,1,2, 3].

Another method is preventive migration of the data to a "fresh" carrier before the old one fails or becomes obsolete, or, in combination with replication, after failure from another copy [11,4]. Essential to this method is also early detection of errors. The latter is the focus of this work.

The most important kinds of external events are fire, floods and earthquakes and violence such as wars and terror attacks. E.g. the recent flood in Prague destroyed besides others a complete library. Even though insurance companies have extraordinarily good data about the first three categories, no study about digital preservation we know about has modeled this risk. We model it as a kind of failure events.

Human and software errors are difficult to grasp. Suitable control mechanism may reduce the risk [6], but in the end, only replication, error detection and migration can help. The here presented method can take into account such factors, if there are empirical risk values available.

Risk analysis is a rather old field and heavily used in technology. In particular air and space industry and nuclear technology have very high reliability demands and have fostered enough research. The other driving force is the insurance industry. The typical methods are fault tree analysis and simulation methods. Simulation methods do not easily allow for understanding the way specific parameters influence the result, and are unreliable for very low probabilities. Under the analytical methods, the continuous-time Markov chains provide a better way to deal with complex transitions than fault trees, used e.g. to analyze the behavior of the RAID systems implementing automatic failure detection and disk replacement [5, 9,18, 20].

Reliability analysis of technological components concentrates typically on the life-time of one system. In contrast to that, for Digital Preservation we are interested in the accumulated risk of many subsequent system life-cycles. On the other hand, replication configurations for digital files are simple compared to a car or an aircraft. This suggests the use of analytical methods. The only respective study for digital preservation we are aware of uses simulation methods [6].

Reliability modelling for long term digital preservation

In this chapter we present analytical modeling techniques to study the reliability of system configurations for digital preservation. The modeling techniques are presented through concrete examples of case studies. The examples are particularly simple, but arguments are made that they describe the most dominant factors. More complex

cases can either be easily added by using different constants or simple extensions of the states investigated.

We start with only looking at the fate of a single file in the system. Different copies may exist on two, three or more disks. For simplicity, we only regard complete media failures. There are two strategies: Replace media after failure, or replace media at regular intervals to prevent failure.

The aging of media and systems is described by the famous Weibull function, with a juvenile phase of decreasing failure rate, a middle phase of constant failure rate, and an end phase of increasing failure rate. Media, that are regularly replaced can be seen as a "black box", which exhibit an average constant failure rate over many replacement cycles, because at equal intervals it starts again with the same juvenile phase. The same argument holds for media that are replaced only after failure: Their failure rate can be approximated by an average constant failure rate, if one looks at many replacement cycles. In both cases, this average can be found for each media in isolation from their Weibull function.

In the sequence, we analyze the critical impact of the time spent for repair, the time spent for failure detection, external events and increasing the number of copies. Errors during copying of media can be described in terms of prolonged repair time, given the due validity checks are executed.

In such configurations, it makes no difference if a single file is lost or a whole database. Always the whole disk fails. Therefore we study in the end the impact of two different strategies: Firstly, a very large database may be distributed over multiple disks, increasing the risk of complete loss significantly. Opposite, the data are split into packages stored separately, decreasing the risk of complete loss significantly, but increasing the risk of partial loss.

Case Study: Simple system configuration with mirrored disks

A collection of digital files (documents). In this case study we assume that content and the number of files are not changing over the time. Storage medium: hard disk drives from the same manufacturer. Preservation policy: for each digital file residing in a hard disk we create a second replica in a second hard disk drive. The disk repair consists of the replacement of a failed disk by a new disk and the data reconstruction using the replicas of each digital file retained in the mirrored disk. The mean time to repair (MTTR) of the disk is assumed to be exponentially distributed, as well as the mean time to failure detection (MTTFD) and the mean time to failure (MTTF) of each disk.

In order to study the reliability of this simple mirror-disk configuration we use a Markov chain model [18, 22, 23] for one pair of mirrored disks (see fig. 1).

In state {2} we assume that both disks are functioning properly. With rate $2*\lambda$ we have a state transition from state {2} to state {1} that indicates that one disk has failed. We assume that disk failures are independent. In state {1} with rate θ we have a transition to state {1D} where the system failure is detected and a repair can be initiated. While the process stays in state {1} it is also possible to have a transition to state {F} where the second disk with the replicas of the documents has also failed and

no repair is possible. State {F} is an absorbing state since there are no transitions out of this state. The process stays in state {F} once entered.

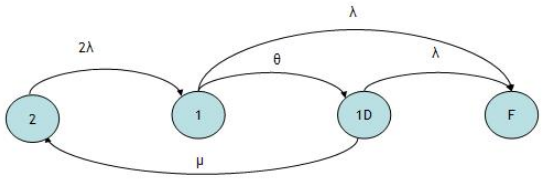

Fig. 1. : Markov model of mirrored disk example

Numerical results. We conducted several experiments in order to study the reliability of this simple mirror-disk pair configuration. As a realistic example, we use $MTTF_{disk} = 1/\lambda = 3\ yrs$,
$MTTR_{disk} = 1/\mu = 50\,hrs$
$MTTFD_{disk} = 1/\theta = 14\ days$.

The resulting mean time to failure ($MTTF_{config}$) is 106,46 years, corresponding to 90% reliability in 11 years (100000 hrs). If both, MTTFD and MTTR tend to zero, $MTTF_{config}$ assympotically tends to infinity (without prove), whereas MTTR+MTTFD limit ultimately the reliability of the system, independent of the individual disk reliability (see fig.2). This leads to the surprising result that investment in quick failure detection and repair must balance the investment in disk reliability. With the given data, a quality increase of the disks can easily be destroyed by lazy error detection (see fig.3).

Extending the mirrored disk example: Adding another backup system

In this section we extend the simple configuration adding another replica for each digital file on a more reliable data carrier (magnetic tape). Fig. 4 shows the Markov model used to describe the system configuration. The disk repair consists of the replacement of a failed disk by a new disk and data reconstruction consists of copying the data from the other working disk with mean $1/\mu_1$ or from the magnetic tape with mean $1/\mu_2$ .depending on the state a disk repair is initialized. States are symbolized by "*n,m*" where n is the state of the mirror disks as in fig.1, and m the state of the backup tape respectively. F marks the complete failure.

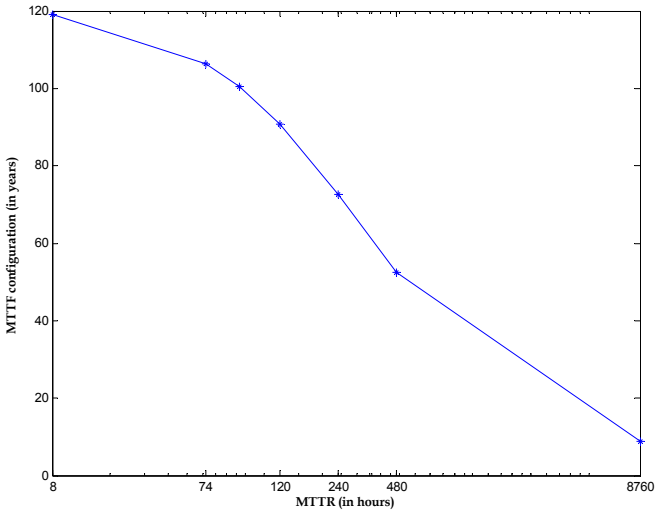

Fig. 2. MTTF of system vs. MTTR disk. This plot shows that mean time to repair does not significantly affect the MTTF of the configuration when the time required to repair a disk is short (<74 hrs). However, increasing the mean time to repair beyond the 74 hours drops sharply the MTTF of the configuration, which leads to the conclusion that MTTR is a dominant factor.

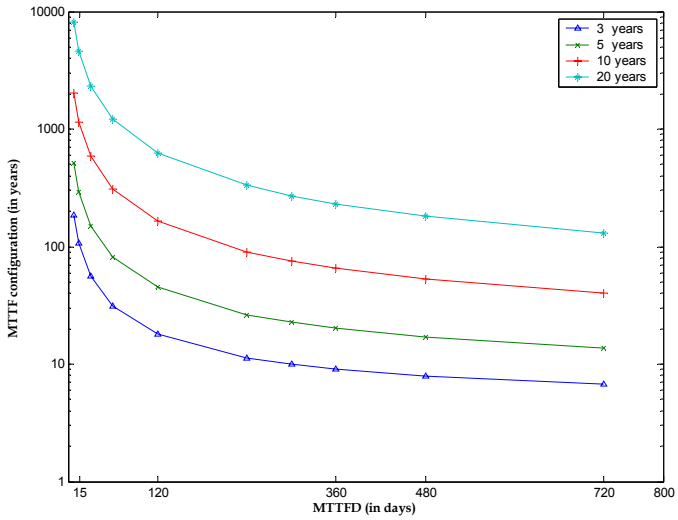

Fig. 3. MTTF of system vs. MTTFD disk. This configuration is equivalent to the RAID I architecture. A RAID system may achieve MTTR+MTTFD of about 2 hours, however it suffers in addition from synchronous failure of both systems which lowers the actual reliability.

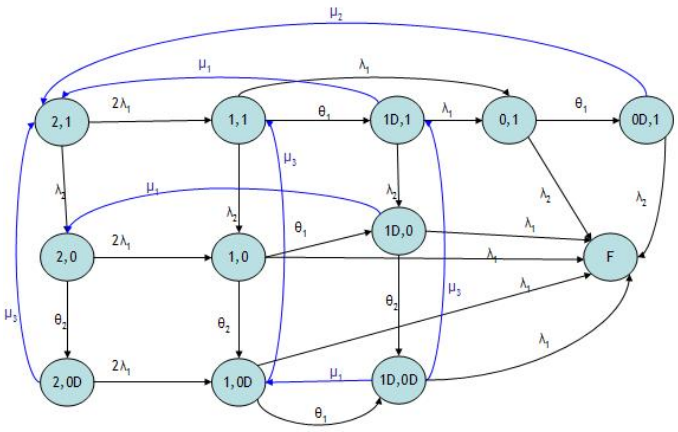

Fig. 4. Markov model of system that maintains three copies for each data file

The input parameters and results are summarized in table 1. The assumption, that a tape is checked every 60 days might be too demanding. Anyhow, we are approaching the interesting reliability area: After 1000 years there is a 32% probability of having lost all data.

Evaluation summary	
For one pair of mirrored disks $MTTF_{disk} = 1/\lambda_1 = 3\ yrs$ $MTTF_{tape} = 1/\lambda_2 = 5\ yrs$ $MTTR_1 = 1/\mu_1 = 50\ hrs$ $MTTR_2 = 1/\mu_2 = 100\ hrs$ $MTTR_3 = 1/\mu_3 = 8\ hrs$ $MTTFD_{disk} = 1/\theta_1 = 14\ days$ $MTTFD_{disk} = 1/\theta_2 = 60\ days$	$MTTF_{config} = 2551\ \mathrm{yrs}$ $R(t) = 0.6746,\ t = 8760000\ hrs\ (\sim 1000\ yrs)$

Table 1. Summary of numerical results

If we add however catastrophic external events such as fire in the physical storage environment of data carrier, the reliability drops dramatically. We investigate two situations:

1. Data carriers (disk and magnetic tape) of the system configuration are stored in the same physical space environment

2. Data carriers of the system configuration are stored in different physical environments. For simplicity we assume that the mirrored disks are stored separately from magnetic tapes containing exact the same set of digital files.

The first case can be modeled by adding a simple fault tree model to the results from Fig.4 and Table 1: the logical combination of events that lead to a system failure (data loss), which are either a fire event in the physical storage place or internal system failures (disk and magnetic tape failures).

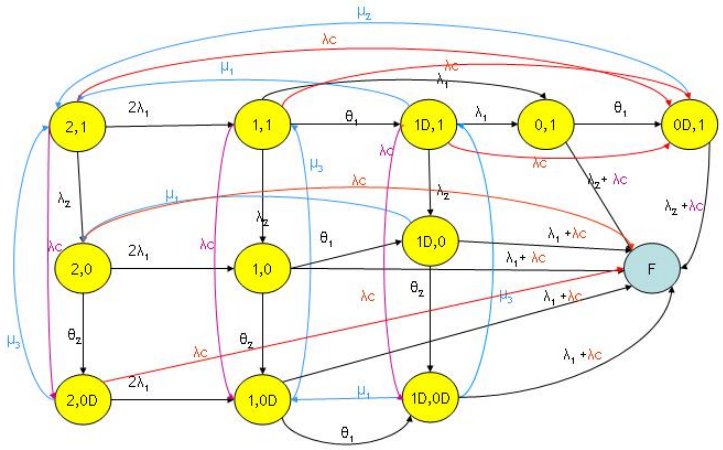

Fig. 5. Markov model with fire event

The second case is described by the Markov model shown in figure 5. According to the information obtained from the National Fire Protection Association the frequency of major fires in residential properties in the United Stated over a 10-year period is 0.9% [8]. Using this estimation in our model we obtain the following results:

1. System configuration without assumptions of catastrophic external event: $MTTF = 2551{,}29$ *years*
2. System configuration assuming one catastrophic fire event and storage of data carriers holding same data on the same physical environment: $MTTF = 773$ *years*
3. System configuration assuming one catastrophic fire event and storage of data carriers holding same data in different physical environment: $MTTF = 2375$ *years*

These results show clearly the need to use different storage spaces for backup copies. Further more, we are still far from the reliability goal set out in the beginning. The use of a fourth back-up copy seems to be inevitable.

Increasing components in a digital preservation system: How reliable does the system configuration remain in the longer term?

If a monolithic database becomes larger and larger, one may distribute the data to more and more mirrored disks and magnetic tapes. Assuming that the system fails when of one of its components fails results in a sharply drop of the system MTTF as shown in figure 6 under the above assumptions with separate backup area.

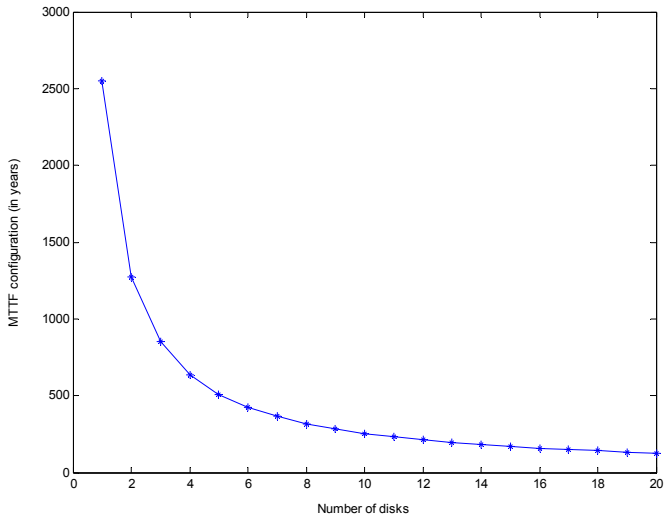

Fig. 6. MTTF system vs. number of components

This means that distribution should be always such that failure of one component may not affect the other. If, e.g., different tables of indices of a database as a whole reside on different carriers, we have this undesirable effect. If, on the other side, the data are distributed into independent clusters, the question arises, how much of these clusters will exist after 1000 years.

We have therefore calculated with a binomial model two cases keeping all other parameters the same: the probability Rk/N that 50% or more of my data exist, and the probability that more than 90% of my data exist, depending on the number of clusters N. Since the clusters are discrete, 90% or more for two clusters means actually that two out of two must exist (K/N = 2/2).

The probability for one cluster to exist after 1000 years is over 60%,, i.e. the probability to lose a monolithic database completely is over 30%. The probability to loose half of the data is still high with 10 clusters, but becomes negligible at 500 clusters. On the other side, increasing the clusters makes sure we loose some data. This surprising but obvious result seems to suggest a strategy, where data are kept in parallel together, and in clusters.

N	K/N (50%)	Rk/N	K/N (>90%)	Rk/N
2	1/2	0.8597	2/2	0.3911
10	5/10	0.8731	9/10	0.0639796
100	50/100	0.9960	90/100	$5.17553 \cdot 10^{-10}$
500	250/500	1	450/500	$2.52509 \cdot 10^{-44}$
1000	500/1000	1	900/1000	$2.00336 \cdot 10^{-63}$

Table 2. Binomial model results for different number of components

Conclusions and Future Work

To our knowledge, this is the first work of risk analysis for a typical museum or digital library configuration using an analytical model. The first results presented here already indicate a high relevance for planning digital preservation. We envisage continuing this work with real life cases and actual data, hopefully leading to a system that database maintainers can handle.

On the other side, the analytical form of the model allows introducing cost models for different configurations, failure detection strategies, hot support lines for repair, data carrier reliability, and additional backup media and storage spaces, so that the optimal cost of a politically set reliability goal can objectively be determined.

References

[1] AES Standard for audio preservation and restoration-Method for estimating life expectancy of compact discs (CD-ROM), based on effects of temperature and relative humidity

[2] Adrian Brown, Digital Preservation Guidance Note 3: Care, handling, and storage of removable media, June 2003

[3] John Van Bogart, Magnetic Tape Storage and Handling, A Guide for Libraries and Archives, National Media Laboratory, June 1995

[4] Brian Cooper, Arturo Crespo, and Hector Garcia-Molina. Implementing a reliable digital object archive, 1999. Submitted for publication to ACM DL 2000.

[5] Peter M.Chen, Edward K.Lee, Garth A.Gibson, Randy H.Katz, David A.Patterson. *RAID: High-Performance, Reliable Secondary Storage*. Submitted to ACM Computing Surveys, October 1993

[6] Arturo Crespo, Archival Repositories for Digital Libraries, PhD Dissertation, University of Stanford, March 2003

[7] Digital Preservation Coalition, HANDBOOK http://www.dpconline.org

[8] W.E. (Em) Fluhr, Paul C. Thomas, M. Arch: Major Structural Damage in Residential Properties

[9] Gibson, G.A and Patterson, D.A. Designing disk arrays for high data reliability. Tech.Rep CMU-CS-92-130, Carnie Mellon University, April 1992.

[10] Den Haag. Testbed Digital Bewaring Migration: Context and Current Status (2001)

[11] Joseph Halpern and Carl Lagoze. The Computing Research Repository: Promoting the rapid dissemination and archiving of computer science research. InProceedings of the Fourth ACM International Conference on Digital Libraries,August 1999.

[12] David I.Heimann, Nitin Mittal and Kishor S.Trivedi. Availability and Reliability Modeling for Computer Systems. In M.Yotis, editor, *Advances in Computer Systems,* volume 31, pages 176-233. Academic Press, San Diego, CA, 1990

[13] H.Harris. Remembering 10,000 Years of History: The Oral History of the Peoples of the Northwest and Related Archaeological and Paleoenvironmental Evidence, 1997

[14] C.Hirel, R.Sahner, X.Zhang, K.Trivedi, Reliability and Performability Modeling using SHARPE 2000. Center for Advances Computing and Communication Department of Electrical and Computer Engineering Duke University, Durham.

[15] Gregory W.Lawrence, William R.Kehoe, Oya Y.Rieger, William H. Walters, Anne R. Kenney. Risk Management of Digital Information: A File Format Investigation (2000)

[16] Raymond A. Lorie. Long Term Preservation of Digital Information. *Joint Conference on Digital Libraries*, ACM/IEEE, June 2001

[17] Manish Malhotra, Kishor S.Trivedi. Reliability Analysis of Redundant Arrays of Inexpensive Disks. *Journal of Parallel and Distributed Computing* 17, 146-151, Academic Press, 1993.

[18] Jogesh K.Muppala, Ricardo M.Fricks and Kishor S.Trivedi. Techniques for System Dependability Evaluation. *Computational Probability*, pp.445-480, Kluwer Academic Publisher, The Netherlands, 2000

[19] Patterson D.A, Gibson G and Katz R. A case for redundant array of inexpensive disks (RAID). *Proc of ACM SIGMOD.* Chicago, IL, June 1988.

[20] Preserving Digital Information: Report on the Task Force on Archiving of Digital Information (1996)

[21] Koichi Sadashige, National Media Laboratory, Host to the National Technology Alliance. Data Storage Assessment – 2002 Projections through 2010. March 2003

[22] Robin A. Sahner, Kishor S.Trivedi, Antonio Puliafito. Performance and Reliability Analysis of Computer Systems: *An Example-Based Approach Using the SHARPE Software Package*. Kluwer Academic Publisher, 1996

[23] Kishor S.Trivedi. Probability and Statistics with Reliability, Queuing and Computer Science Applications, Wiley-Interscience Publication, 2002