Storage Strategy Tools

Matthew Addis, Mariusz Jacyno, Martin Hall-May IT Innovation Centre,

University of Southampton, UK

Richard Wright Archive Research,

BBC Research and Development, UK

Introduction

This article describes the challenges of selecting a storage strategy for long-term retention and access of digital content and how some of the tools developed in the PrestoPRIME[1] project are designed to help.

Long term storage of digital content can be a significant cost, although not typically the biggest cost factor for preservation and access – digitisation, cataloguing and rights management often dominate for many audiovisual archives. In many cases, storage is also a cost that falls rapidly, with some claiming that storage is essentially becoming ‘free’, at least in comparison with other costs. However, currently, storage is expensive enough to be a significant part of preservation budgets and as such requires due attention. Storage is also a recurring cost. Worse still, in the case of digital content stored on IT systems, storage is a recurring cost where payment has to be kept up or otherwise content is rapidly at risk of being irretrievably lost due to obsolescence and storage failures. This is where the real ‘cost’ lies: the cost of doing nothing or the cost of doing it wrongly – both result in loss of content, i.e. the ‘cost of loss’.

Storage is not about giving digital content a safe harbour where it can live out its days. Given the ever increasing drivers for making archive content rapidly and easily accessible, and with a ‘holy grail’ of having ‘everything online’ with ‘instant access’, storage should be considered as a fundamental component of enabling access as well as safe keeping. For many audiovisual archives the reason content is kept is so it can be reused, repurposed and enjoyed again. The type of storage used to store content has a big impact on the accessibility of content being stored: compare film in a freezer with disk servers on a high-speed network. Yet the types of storage that provide the fastest access are not necessarily the ones that have the lowest cost of ownership, nor the highest degree of safety for the content within them. You can strive for low-cost, high-safety, and fast-access – but it’s not possible to achieve all three at once!

Here lies the dilemma. There is a trade-off between cost, ease of access, and safety of digital content when using any type of storage. There is no ‘one size fits all’ solution. Neither is there a ‘one size fits all’ set of needs from archives. The challenge is one of combining different types of storage and matching the combined characteristics to an individual archive’s needs. This means matching storage strategies to the needs of an archive both today and in the long-term. We attempt to tackle this challenge with our simulation and modelling tools, in particular by allowing the costs and risks to be compared – which we call the ‘cost of risk of loss’ approach.

[1] PrestoPRIME http://www.prestoprime.org/

State of the nation

There is an increasing body of work on the total lifetime costs of storage [1][2][3]. Costs include power, space, cooling, maintenance, migration and management [4] in addition to the more obvious costs of media and the physical servers or shelves used to store it. Some of the cost elements fall rapidly over time, e.g. following Kryder's law[2], although there is concern on how much longer this will hold[3]. Some costs, for example power, may actually increase, although the net effect is still downwards when increased storage densities and efficiencies are taken into account. If past performance is an indicator of the future, then the Total Cost of Ownership (TCO) of storage will continue to fall for at least the next decade, and at a significant rate. For example, the Internet Archive can be used to look back at the costs of online storage service providers (e.g. Amazon S3) over time, with the charge per TB per year seen to halve every 2-3 years [1]. This is slower than the rate of fall of underlying media costs, e.g. the cost of a hard drive in a local electronics store halves every 18 months or less. The slower rate of fall represents the effects of all the other cost factors that make up TCO (physical media can be as little as 10%)[4]. So confident are some organisations [5] in this trend of falling costs that that they will accept a one-off payment today in exchange for a commitment to store content indefinitely – a so called 'forever price'.

There is also a significant body of work on the reliability and safety of storage for digital content [6,7], much of it focusing on IT technology, e.g. Hard Disk Drives, data tape and optical discs. This work is well known to the vendors of such storage, as evidenced by increasingly complex systems to guard against failures, e.g. double or triple parity RAID arrays [8] and tape libraries with automated data integrity checking[4]. There is some debate over the value of the metrics used by the storage industry for safety, e.g. as highlighted by the paper "Mean time to meaningless" [9] with better ways to measure safety also proposed - but this only serves to highlight that IT storage systems aren't fundamentally reliable ways to store data in the long-term. That is unless a suitable active data integrity management strategy and set of processes are put in place. This active management approach can now be seen at all levels including archive asset management systems[5], file systems[6] as well as inside storage systems as already discussed.

Storage reliability and safety is faced by the people using these storage systems 'day to day', e.g. IT departments and their systems managers – although this is often learnt the 'hard way' through experience of loss rather than knowing what best practice to apply upfront. The main problem is that this knowledge of storage reliability is relatively unknown to the archive community – generally speaking, good practice has yet to be widely established when using IT systems for digital preservation of AV content. Worse still, techniques commonly used by archivists for storage in the 'analogue domain', e.g. 'putting media items on shelves', are often assumed to translate into the 'digital domain'. In fact the opposite is true. Storage of digital data requires active management and a plan of regular interventions (refreshes, migrations, integrity checks and repairs) to ensure data remains intact and the systems are in good shape.

[2] http://en.wikipedia.org/wiki/Kryder%27s_Law#Kryder.27s_Law

[3] http://www.storagenewsletter.com/news/disk/hdd-technology-trends-ibm

[4] Examples in 2011 include SpectraLogic and Quantum adding this feature to their libraries/software stack

[5] For example, DIVArchive from Front Porch Digital

[6] For example, ZFS. http://en.wikipedia.org/wiki/ZFS

Whilst information exists separately on both the costs and risks of different types of storage, a lot less work has been done to combine the two, especially in the context of storage being an enabler for access to content. Existing work also tends to focus on single storage systems rather than their combination when adopting good preservation practice of ‘multiple copies in multiple locations, using diverse technologies’. We attempt to address this gap in our work.

Cost, access and safety don’t go hand-in-hand

There are many different approaches to storing digital data. Each has its own strengths and weaknesses in terms of safety, ease of access, lifetime and cost. For example, long lived storage technology includes printing bits to film[7] but costs per TB are high and access requires a film scanner. IT storage technology includes data tape or HDD, with data tape offering lower TCO at high data volumes and a good level of safety, but not the rapid ‘random access’ that HDD servers can afford - but for HDD servers this increased ease of access comes at the expense of increased cost and shorter time to obsolescence. Making more copies of content is of course one strategy for increasing safety, e.g. an extra copy in an offsite deep archive, which is essential to guard against some catastrophic types of failure, but does so with significant extra cost. The content itself can be encoded to make it more resilient to failures in storage [10], but this adds complexity and potentially a different set of risks from using a format that is not ‘standard’ and hence itself at risk of becoming obsolete. Error correction is commonplace in almost all types of digital storage and works tirelessly behind the scenes to correct a multitude of sins. Further layers of error detection and correction can be added to storage to catch and repair problems (e.g. the use of checksums and ‘scrubbing’), but this adds additional cost. Errors can also be concealed. When it comes to errors, it is not actually lost bits and bytes that typically matter - rather the impact they have on the ability to use the content. Here it is possible to conceal this impact, e.g. artefacts introduced by block-level corruption in a video stream can be interpolated from adjacent frames and often made visually acceptable.

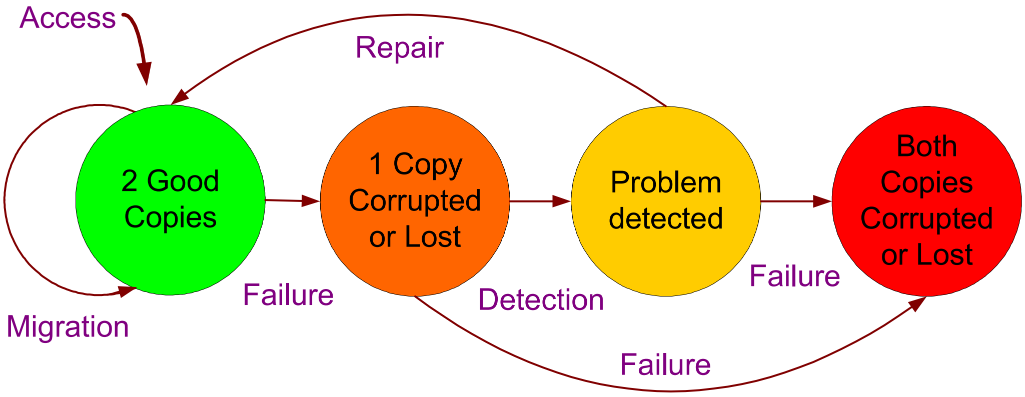


Figure 1

Figure 1 above presents a conceptual model for comparing and combining different storage approaches. With reference to the diagram, the bedrock of data safety is to keep multiple copies of content (green circle), typically by using different storage technologies and in different locations,

[7] For example, DOTS from Cinevation. http://cinevation.net/

and ideally operated by different people. This guards against major risks (e.g. it allows disaster recovery), but it also guards against unanticipated problems with individual technologies and processes (i.e. it ensures that the eggs are not “all in one basket” at any level). For each copy, there is the need to regularly migrate each component of the technology stack (hardware, operating system, management software, file formats, etc.). One or more of the copies is also used to serve access requests. However, there is always the chance that one of the copies could be damaged or lost due to some form of failure in the system (orange circle). Only after this problem is detected (yellow circle) can any action be taken to repair, replace or conceal defects in the damaged or lost copy. If at any time before this process is complete something happens to the second - and only remaining good copy - then there is a risk that both copies could be permanently lost or damaged (red), and the content could be lost. Depending on the type or damage or loss, there is the option of doing partial repairs, concealing defects (a common approach in digital video), or reconstructing a good copy from parts taken from two damaged copies. All these options have additional costs and benefits. The rate at which transitions happen between the states dictates the length of time during which content is at risk. Every transition also has a cost. Therefore, whilst superficially quite simplistic, the model above provides a framework for the total cost and total risk to be assessed and storage strategies to be compared. A detailed qualitative comparison of a wide range of storage techniques is available in an article by the authors in the SMPTE digital imaging journal [10], also available online in extended form as a public PrestoPRIME report [11].

Tools

Our approach to quantitative comparison of the costs and risks when using storage starts with a simple but flexible storage model (Figure 2) that consists of a set of storage systems (e.g. tape library, HDD server). Each storage system has the function of accepting files for storage (writes), returning files from storage (reads) and storing the data inside the files using some form of physical media (hard drive, data tape, optical disc etc.). The model includes a ‘controller’ that mediates access to the underlying storage. The controller could be a person, i.e. manual operation of ‘media on shelves’ archive, or the controller could be automatic, e.g. storage management software operating a tape library.

The reason for choosing a simple model is so it can be applied to both automated hardware/software, e.g. a HDD server, and to more manual processes, e.g. data tapes on shelves where archive personnel put new tapes onto shelves and retrieve existing tapes to serve user access requests. When writing or reading files, various operations may be applied, e.g. encoding or applying error correction. Depending on the system being modelled, this could be by firmware on a HDD, the RAID controller in a HDD array, integrity management in a ZFS filesystem, manual integrity verification by an operator, or a combination of all of these. Likewise, various failures or errors could occur, both latent (failures that happen silently [12]) or extant (failures that are immediately detected). Failures can range from ‘bit rot’ in a HDD system through to accidental damage from manual handling of discrete media, e.g. data tapes. These failures can happen (a) when data is written, (b) when data is read, and (c) when the data on the physical media is in effect ‘doing nothing’. In our model, these are all represented through error rates for read/write/store actions. The actions each have a cost, which forms one part of the associated cost model (e.g. one-off ingest cost per file when adding it to a storage system, access cost per file incurred each and every time it is retrieved from the storage system, and the on-going storage cost per file when it is inside the storage system).

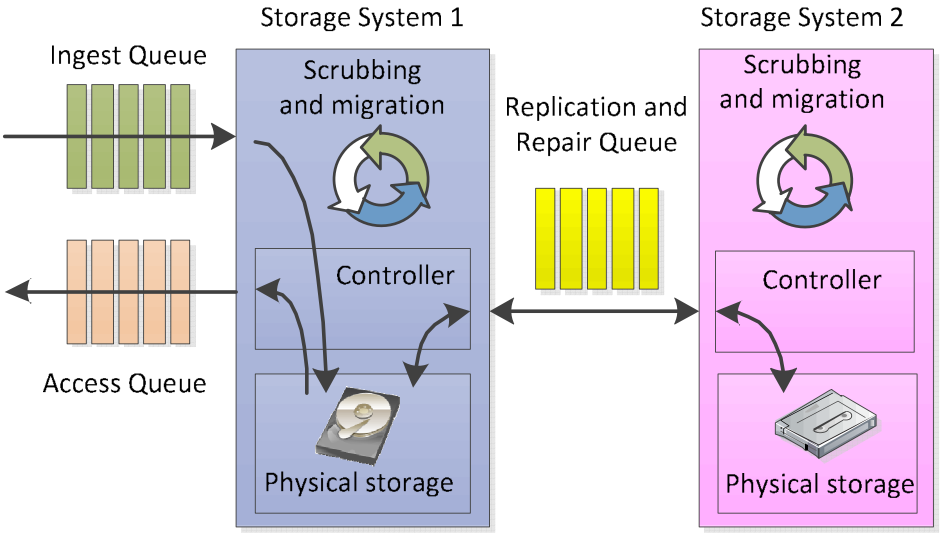


Figure 2 Archive storage strategy

In the model, one or more storage systems are then combined into a storage strategy. Figure 2 shows just one example that combines two storage systems. The storage strategy includes ingest for new file arrivals and access queues to serve user requests for content. The strategy determines how files are allocated to storage, how they are replicated, and how they are repaired if there are failures. Resources can be allocated to serving ingest, access and copy operations as well as for activities within each storage system, e.g. integrity checking and repair. It is often these resources that are limited, especially at peaks of workload, so limits can be set in the model to allow investigation of what happens when resources run out.

We have made two tools available[8] that allow projections of cost and loss over time. The first tool takes a very simple approach that allows storage systems to be characterised in just 4 parameters (cost of storage, cost of access, latent corruption rate, access corruption rate). Two storage systems are combined into a storage strategy (2 copy model). This includes parameters for the frequency of checking and repair of the data (scrubbing), the number of files to start with and how fast this increases, how often access to these files takes place, when migration is needed and how costs change over time. The tool comes pre-loaded with parameters based on real-world storage systems, with some examples given in Table 1 for a scenario of storing 25GB files (e.g. approx. 1 hour of SD video at 50Mbit/sec). Full derivation of these values can be found in [11] and they of course vary over time and with the size of the files.

	Latent corruption rate (files)	Access corruption rate (files)	Storage costs (Euro per GB per year)	Access costs (Euro per GB)
HDD online (server)	1 in 750	1 in 500	1	0.1
HDD offline (shelves)	1 in 100	1 in 500	0.1	5
Data tape online (library)	1 in 100,000	1 in 10,000	0.5	0.1
Data tape offline (shelves)	1 in 10,000	1 in 1,000	0.05	5

[8] http://prestoprime.it-innovation.soton.ac.uk/

Table 1 example parameters for storage of 25GB files

The simulation then projects the number of files that are ‘alive’, ‘at risk’ or ‘lost’ over time (lost means that both copies have become corrupted, at risk means that one copy has become corrupted, alive means that there is at least one copy that hasn’t been corrupted). An example is shown in Figure 3 for an archive that stores its contents using HDD on shelves. For simplicity, the number of files in the archive remains constant and there is no regular ‘scrubbing’ to check integrity (due to the costs of doing this for a HDD on shelves model). The number of files at risk climbs year on year, as does the annual number of files lost, until a migration point which provides the opportunity to detect and repair losses.

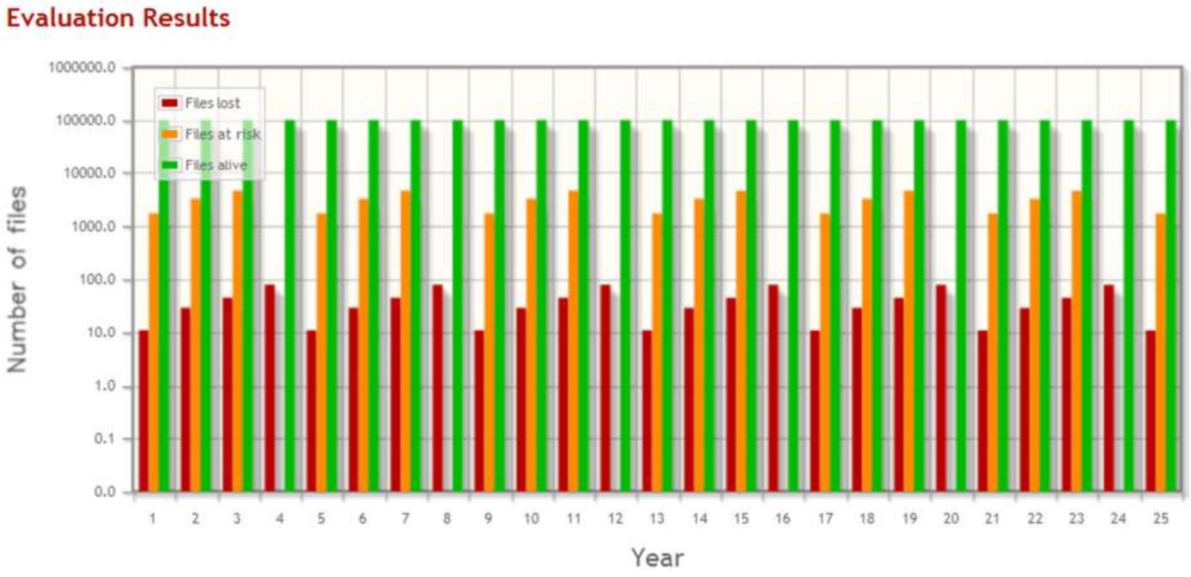


Figure 3 Files at risk and lost for a 25 year projection for a HDD on shelves strategy

The corresponding costs are shown in Figure 4. As with the risk/loss projection, cost per year is fixed to make the graphs easier to understand for the purposes of this paper. The cost is broken down into each storage system (with one higher than the other due to user access to content being delivered through storage system A and not system B). The spikes in cost every 4 years correspond to migrations. It is then possible to use the tool to adjust parameters and look at the consequent impact on costs and loss over time. In this way, different storage strategies, e.g. tapes and HDD in servers/libraries can be compared with a ‘media on shelves’ approach.

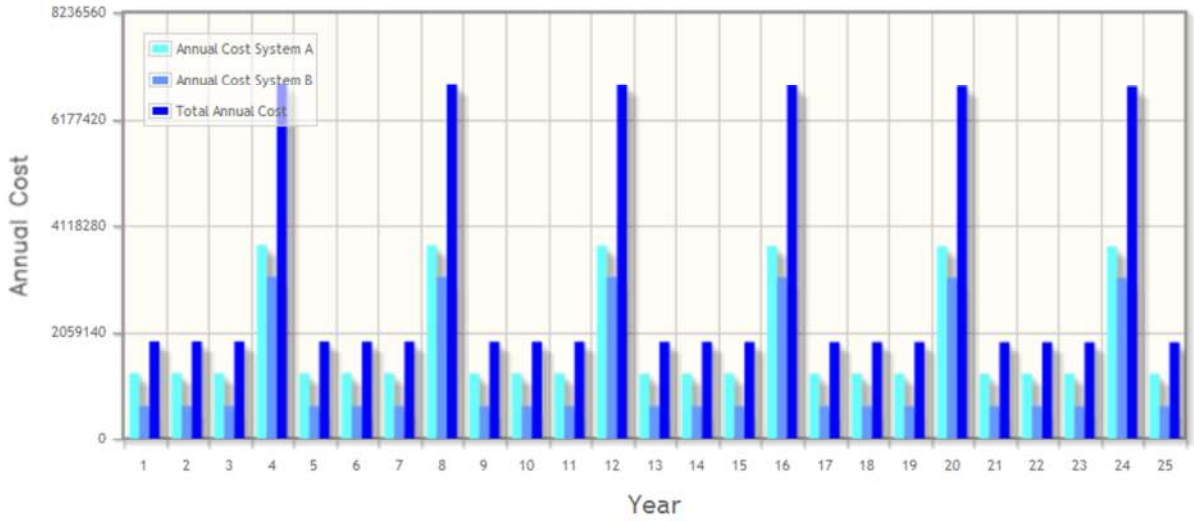
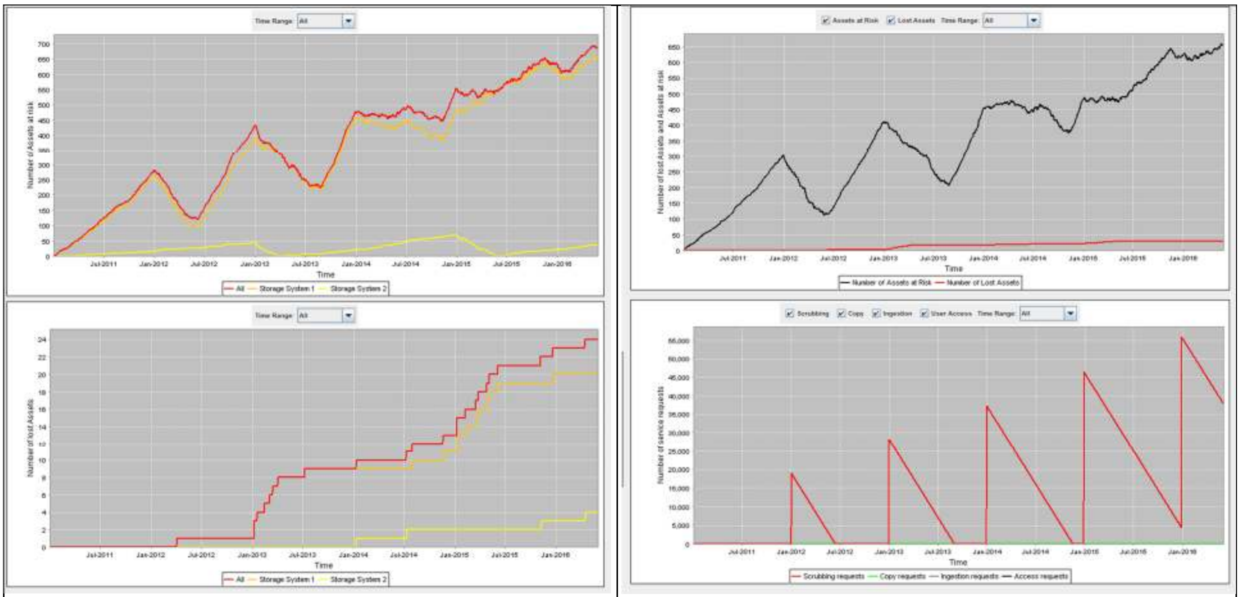


Figure 4 Cost for a 25 year projection

It should be clear that the projections from the simple model are exactly that – simplistic. They do not provide the detail or accuracy for use in business planning or day to day operation. The objective is however only to provide a simple comparison of some of the strategies that could be used for storage. In this respect, the tool is a useful educational aid in illustrating the importance of an active storage management approach if data integrity is desired over the long-term.

To allow a more in-depth analysis of storage strategies, the second tool developed provides an interactive simulation tool based on a discrete event simulation approach. During the simulation, time ticks away (e.g. 1 second of the simulation might correspond to 1 week in the real world) and events are generated (e.g. corruption of files in a storage system, requests to access a file, new files to be added to the archive). These events then trigger actions, e.g. a copy/repair process, which is then added to the queues of the storage systems involved. A storage system processes items in its queues according to how much resource it has available (e.g. serving access requests sequentially or in parallel). The available capacity of the resources used by each service determines how many items are processed for each tick of the clock, and at what cost.

The user can interact with the simulation as it progresses, e.g. changing the amount of resources available or changing the policy for data safety (e.g. making more copies or checking them more often). In this way, the user is in effect playing a game that helps them understand how to react to and manage events that they might see in practice when operating a real system. For example, there is also an option to simulate 'disaster scenarios': rare but catastrophic events where large fractions of the storage become temporarily or permanently unavailable.



Overall risk and loss profile

Risk and resources in storage system1

Figure 5 Interactive simulation of a combined HDD and data tape storage strategy

An example of a simulation is shown in Figure 5, again for a 2 copy model, but this time modelling a combination of HDDs in a server and tapes in a library. Periodic scrubbing is included as well as migration, and there is growth in archive content over time with a proportionate amount of access requests to go with it. All starts well with a low rate of content loss and regular integrity checking well within system capacity. However, as the archive grows there is an increasing access burden and more content to manage. Priority is given to ingest/access over background scrubbing with the result that integrity management starts to suffer and eventually runs out of resources with a consequent increase in loss of content. At any point during the simulation it is possible for the user to adjust the resources in the system in order to counteract this effect, and hence estimate how the capacity of the system will need to be extended over time. The example above is intended to be illustrative and the tool allows much more sophisticated simulations to be run, recorded and rerun.

Repeated running of the simulation with variation in input parameters allows a ‘map’ to be created of the cost/risk ‘landscape’ for a given storage strategy.

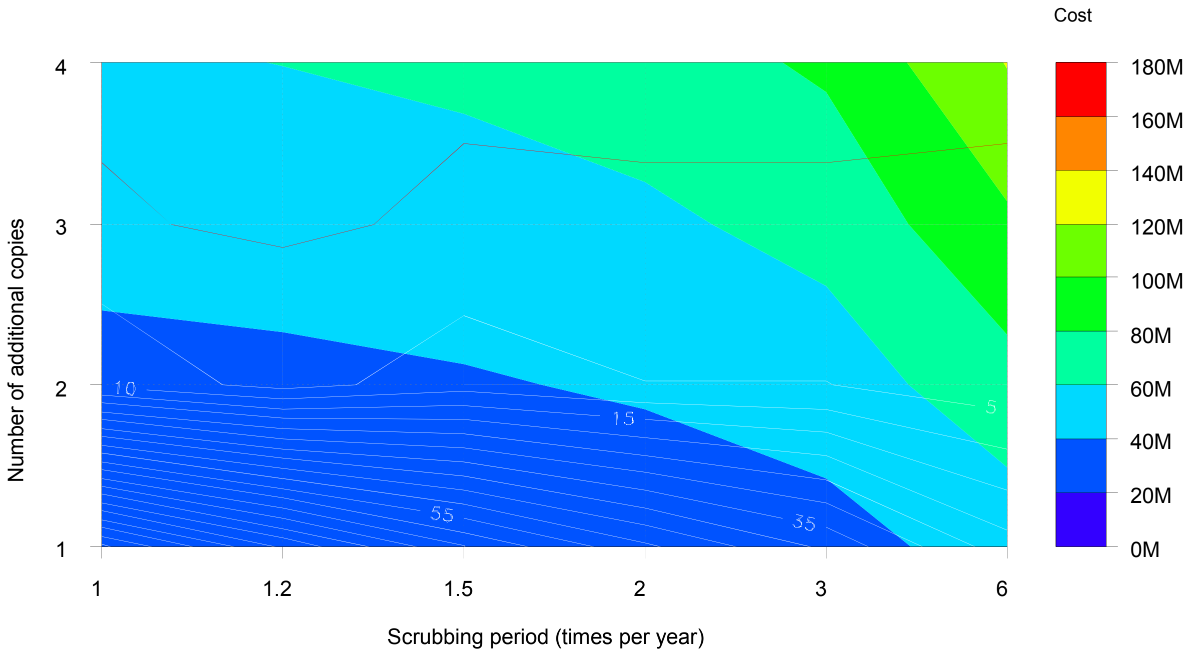


Figure 6 Cost of risk map

An example map is shown in

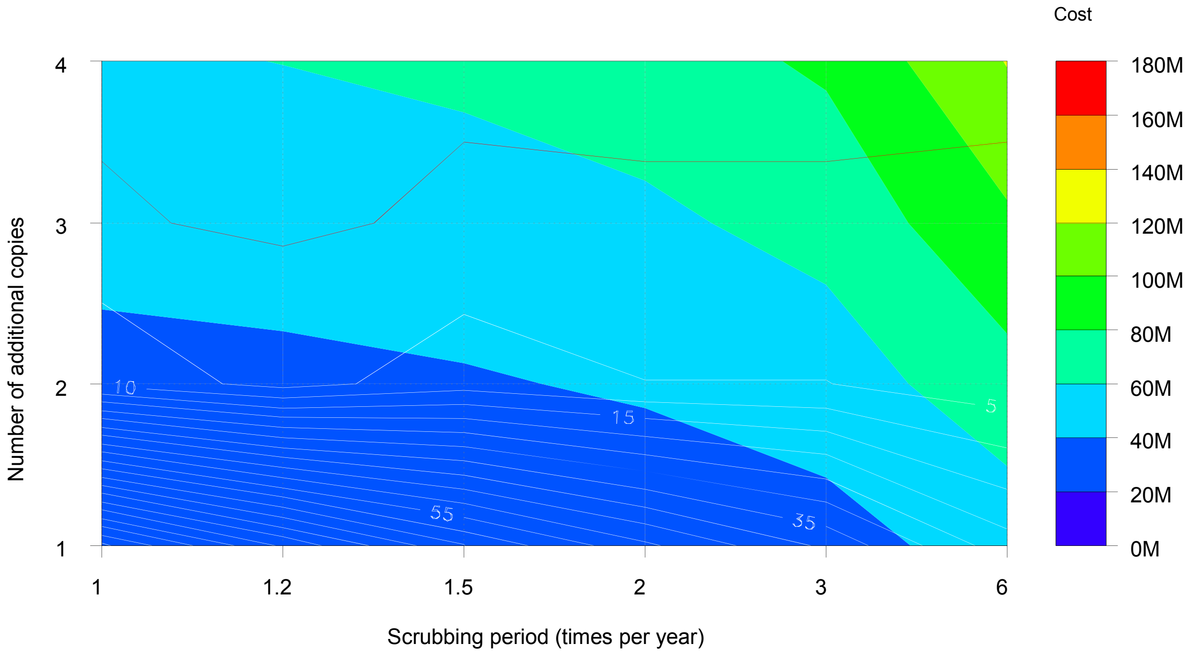


Figure 6. This shows a single storage system where the number of additional copies of a file stored in that system and frequency of integrity checking (scrubbing) both impact on both the cost and the risk of file loss. The boundary between adjacent coloured bands represent configurations of equal cost. The white contour lines are lines of equal risk of loss and represent the peak number of assets where only one uncorrupted copy is left (which is typically a transient position because scrubbing picks up and repairs the corruption). The red contour line is the boundary between the probability being in favour of at least one file loss or no file loss over a 10 year period. Increasing the number of copies reduces the risk of loss, as expected, but also increases the cost because more storage

capacity is needed. Increasing the scrubbing frequency also reduces the risk of loss, but again increases cost because of increased access to data and equipment needed to compute checksums for a larger volume of data. For the particular parameters used for the storage system shown, it is more cost effective to add more copies than it is to scrub them more often and there is a 'sweet spot' that balances the two to give the lowest cost of zero probable loss over 10 years. Different storage strategies have different balances, which is why a tool to allow the trade-offs to be analysed on a case by case basis is so important. If zero loss is required then it also allows the area on the landscape to be found that is sufficiently 'far' from risk contours to give some margin for error between the simulation, e.g. because of lack of precise input data, and what happens in the real world, e.g. because things rarely go according to plan!

Next steps

The tools described in this paper are still in their early days. There is much that could be done to add functionality and improve them. The next steps of our work are three-fold. Firstly, we are investigating how we can best validate the models, i.e. give confidence in their results. This can be achieved in several ways, e.g. (a) comparison with analytical approaches for simple tests cases that are also tractable this way, (b) comparison with the findings of field studies and real archive experience – although this is hard because whilst some data exists it is insufficient to validate all aspects of the model, and (c) validation of the implementation by allowing independent inspection of the source code and design of the model to pick-up potential bugs – which is one reason why the tool is available as open-source and we are transparent with our design assumptions and testing. Secondly, we are planning to develop support in the model for higher-level preservation functions, e.g. transcode of file formats (another type of migration) and repair/concealment of the impact of corruption at the AV level, e.g. mapping the effect of data corruption onto different video formats and how the corruption could be concealed or repaired. Thirdly we are looking at ways of making the tools simpler to use. The simpler the tool, the less time and investment it takes to learn how to use it and get the required results. However, if this is achieved by making the tool too restricted then the model and the real-world diverge and the results of modelling are less valuable. There is a balance to strike between model complexity, model accuracy, ease of use and what users of the tool are willing to invest in terms of time and effort learning how to use it.

Acknowledgements

PrestoPRIME is an EC supported 7th Framework Programme ICT project (FP7-231161) coordinated by INA (Institut National de l'Audiovisuel) in France. Partners include BBC, RAI, ORF, B&G and others. For further information see www.prestoprime.org

References

[1] Matthew Addis and Richard Wright (2010). PrestoPRIME Deliverable D2.1.1 "Audiovisual preservation strategies, data models and value-chains". http://www.prestoprime.org/project/public.en.html

[2] David Rosenthal's Blog. See post "Modelling the Costs of Long Term Storage" posted on 27 September 2011. http://blog.dshr.org/ Consulted 12 October 2011

[3] Richard Wright, Matthew Addis, Rajitha Weerakkody (2010). "Century Store, Real Options, Real Costs". Published at the AMIA/IASA 2010 Joint Conference, 2-6 November, Philadelphia, USA.

[4] Barroso, L. A. and Holze, U. (2009). The Datacenter as a Computer: An introduction to the design of warehouse-scale machines. Google Inc. Synthesis Lectures on Computer Architecture no. 6. Published by Morgan and Claypool.

[5] Serge J. Goldstein, Mark Ratliff (2010). DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data. Office of Information Technology, Princeton University. August 27, 2010.

[6] Elerath, J. (2007). Hard Disk Drives: The Good, the Bad and the Ugly!, Queue 5, 6, p28-37. http://doi.acm.org/10.1145/1317394.1317403

[7] Jiang, W. *et al.* (2008) Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. FAST '08. http://www.usenix.org/events/fast08/tech/jiang.html

[8] Adam Leventhal (2009). Triple Parity RAID and beyond. ACM Queue vol. 7, no. 11. http://queue.acm.org/detail.cfm?id=1670144

[9] Kevin M. Greenan, James S. Plank, Jay J. Wylie (2010). Mean time to meaningless: MTTDL, Markov models, and storage system reliability. In proceedings of The 2nd Workshop on Hot Topics in Storage and File Systems (HotStorage '10), June 22, 2010, Boston MA, USA

[10] Matthew Addis, Richard Wright, Rajitha Weerakkody (2011). "Digital Preservation Strategies: the cost of risk of loss for AV Content". Jan/Feb 2011 edition of the Motion Imaging Journal of the Society of Motion Picture and Television Engineers (SMPTE).

[11] Matthew Addis, Mariusz Jacyno (2010). PrestoPRIME deliverable D2.1.2 Tools for modelling and simulating migration-based preservation. http://www.prestoprime.org/project/public.en.html

[12] David Rosenthal (2010). Keeping bits safe: how hard can it be? ACM Queue vol. 8, no. 10 http://queue.acm.org/detail.cfm?id=1866298