

Trusted Media Storage

Yao Yanjun

Helsinki University of Technology

yyao@cc.hut.fi

Abstract

Until quite recently, most families either did not need or could not afford more than one computer, and home networks were only for technophiles. Things are changing very rapidly, and as a consequence of that, more and more families are employing home networks to build connectivity between their computers and devices which can connect to the internet. Along with the development of home networks, the amount of the digital data and the requirement of personal digital media increases day by day. The purpose of this paper is to identify a list of needs for trusted storage of home networks. In the following sections, an analysis of those needs focusing on data management and security with respect to the existing technologies will be discussed.

KEYWORDS:

trusted storage media, annotation, digital archiving, security, network file systems

1 Introduction

In the past few decades, we have been facing a great change in the gradual digitization of the record of videos. It is a common phenomenon that many families have recorded their family history only using films, videos or digital pictures. Modern devices, such as the digital camera and digital vidi-con, can provide great conveniences to common users, hence it is easy to record huge amounts of digital data whenever we want. People can even use services to convert old format of non-digital records into digital formats. As a consequence of the decreasing cost of storage medium, and improving technologies of video recording and data compression, several families are even considering storing decades of their family histories in digital form on home computers.

Some of that digital data is really important and precious, such as pictures of the children, a special movie of the family get-together or original works of parents, which people might want to review several years later and keep between the family only. In that case, losing any of it would be a great regret. Thus the following concerns arise: what if after many years when people want to read the data, which is not lost, but can not find which directory it was saved in? What if the data that people want to read gets lost because of some unexpected reasons? What if some data that should be kept secret is exposed to the public as the result of the malicious actions of hackers?

In order to diminish the concerns mentioned above, according to the property list provided by OceanStore (a global

persistent data store) [1] and the characteristics of home network digital storage, the needs of the trusted storage are summarized into the following two points: the convenient management of media and a good security of data.

The convenient management comprises two parts, easy to store and easy to read. As the price of storage media is decreasing day by day, and data compression technology is nearly mature, the "easy to store" issue will not be discussed in this paper. The focus is "easy to read". People may find that it is really hard to find some pictures out of a massive number of them, not mention to find a specific scene of one movie. Sometimes, even if we find something, it may not help the owner to remember anything. Nowadays, digital archives are managed through annotations to retrieve information. Hence in the following section, both annotation technology and digital archive systems will be discussed.

On the other hand, good security can be separated into confidentiality, integrity and availability. The trait of trusted media storage we are talking about here is that the data needs to be kept for a very long time span, which makes time itself a threat. In section three it will be described in detail. Commonly, the most important aspect for the user of this home networks is availability. Nowadays several network file systems have been proposed to backup data, for example pStore [5], OceanStore, The Coda file system [6] and PAST [7]. In order to give the reader a precise view of the security issues in trusted media storage, we will focus the discussion on the security related to long-term storage systems in the following sections.

2 Annotation

Annotation, which has a long history in printed documents, is being implemented in the scope of digital data, such as text, images and video. Annotations are typically text or graphics, either in printed or digital media. Some systems also use other media (audio and video) as annotations. The goal of annotation is to provide a method to describe the content for later retrieval. As a consequence of that, annotation provides a way to add more information to existing documents, and it can serve multiple purposes such as to highlight relevant parts of a document, to add notes when the document is presented, and so on.

2.1 Media Annotation

In the media realm, each annotation identifies a media element that would be displayed when the media reaches a

given point in time. Unlike annotations in printed documents, which are spatially distributed, media annotation has to be considered in both temporal and spatial dimensions. Annotation sets are used to group annotations made by users, and the same media can be marked with more than one annotation set. Through annotation, users can make new connections, gather and interpret materials, and promote an accretion of both structure and contents. Media annotation is usually aimed at catalogue, using formal annotations or metadata, both of them particularly useful for search purposes.

Metadata, so called "data about data", can generally be thought of as information that describes or supplements the central data [3]. Metadata can be considered extra or essential. Metadata, which merely adds new information and is not critical to the function of the main data, can be considered as extra data, such as the metadata produced by digital still cameras describe the settings used for the pictures. On the other hand, metadata can be considered essential to the proper functioning of the main product, such as metadata on a Zip disk where it provides information about the write-protected status of the disk.

Because of the context it provides and the ways contextual information can be used, metadata is valuable even when it is not essential to the proper functioning of a product. When data is made available to a potential user, the user (human or computer) must put the data into an existing model of knowledge, and may ask questions to do so. For example, in the case of an image, typical questions include "When was this taken?" and "Who and what are in this image?" Metadata provides context to answer many of these questions. In sophisticated data systems, the metadata, the contextual information surrounding the data are also very sophisticated, in order to answer many questions that can help the users understand the data.

It is necessary to keep track of all metadata created even in the early phases of planning and designing. Attaching metadata only after the production process has been completed is not an economical choice, because if metadata is not created immediately at the time the media is stored, the user may have to spend a lot of time on restoring it manually afterwards. Therefore, it is necessary for different groups of resource producers to cooperate using compatible methods and standards. The metadata should be manipulated, and it must adapt according to the content of the resource and be merged when resources are merged. Not all of these requirements are performed nowadays.

2.2 Digital Archives

As we mentioned before, although it is easy to record and edit huge amounts of digital data nowadays, archiving and retrieving is not an effortless task. As time goes by, the amount of digital data grows, too. Some of us might have the experience that it is really time consuming to find the photos of a certain person in a certain place or time from a huge amount of photos, not to mention from home movies. In order to provide good management of digital data, digital archive comes to our sight. In this section we will introduce some issues related to digital archive and media information retrieval.

A digital archive, which includes the preservation and retrieval of digital data, is basically managed through annotations. Usually, annotations provide a window of access into a digital archive. Even when the perfect storage strategies are employed by the archive systems, the digital archive would still be meaningless without annotations, as the purpose of storage is to retrieve, and if there is barely any way to retrieve, then there is no need to store the data at all. With proper annotations, the archive can be accessed more efficiently, and by managing them, the archive can remain fresh and current.

Interaction, which serves the user to annotate and access archives, is an essential ingredient of archival systems. It is the key between the user and the system. Normally, the archive system may provide basic annotations which usually does not fulfill the requirement of users, so the reannotating process is provided. The user should get the data by information query before reannotating it. There are several types of information queries [4]: query from a controlled vocabulary, query by keywords or descriptors, and query by full text, full audio or full visual examples. In practice, what the user gets from those query types is not the direct answer. The result of the query leads the user to an interactive session with the system, where advanced visualization and relevant feedback from users is iteratively used to bring the user closer to the desired information. Ideally, the role of the system is actively participating in finding the best solution by posing the most informative questions or showing the most informative results to the user.

In the future, digital archive systems will automatically learn from the pattern of interaction of the users and the annotation of the data to provide faster retrievals [3]. The retrieval process works like this. Suppose similar content is likely to have the same annotation, so that after initial annotations are provided by users, the system can select collections of similar items which may have the same annotation. By manually filtering out the small proportion of wrongly labeled items, then the user can make a complete annotation of a collection of items in a shorter time.

The function of interactivity puts strict demands on the compute, storage, and display power of the system. Commonly users want immediate feedback on their queries. However, if external examples are used, computation of a large set of relevant descriptors and comparison of the descriptors of all elements in the dataset are needed. In order to limit the search, advanced database techniques are required. Nevertheless, interactivity can compensate the lack of context of the computer [4]. In a full interaction scheme, not only the query but also what is to be considered similar and what are to be considered as positive or negative examples may be modified. Unfortunately, current content-based retrieval systems still can not fulfill all what is to be expected in the near future.

3 Security of Trusted Media Storage

In contrast to traditional storage systems, in which data lifetimes are measured in months or years, data lifetimes for media storage are measured in decades. As the characteristic of storage systems is write-once and read-maybe, which puts

strict demands on the preservation and retrieval properties of long-term storage systems. What is more, the characteristic of long term brings new security threats to the storage systems.

Roughly, there are two categories of attacks against storage systems, passive ones and active ones. The passive attack is mainly restricted to eavesdropping on communication channels, but the active one would do malicious actions to break in and compromise the data in a storage system. The huge time span makes time an active adversary, which is constantly working against the system and threatening to compromise the information. In this section, we will talk about threats to data in long term storage in the CIA (Confidentiality, Integrity and Availability) Model [8].

3.1 Threats to Confidentiality

Confidentiality means that only authorized entities have or can acquire knowledge of the contents of some data. The presence of long lifetimes introduces some unique threats to confidentiality.

Commonly, storage systems use encryption to provide the secrecy of files. As a consequence of long-lived encryption, data is persistent and the key should be kept maybe forever. The fact is that an encryption strategy that seems perfect today may expose vital weakness tomorrow. Although the user keeps the passwords well, the data would still be compromised. In order to avoid this situation, the data should be re-encrypted in a certain span of time. A lot of situations may lead to re-encrypting the contents of files, such as key rotations, compromised keys, compromised encryption algorithms and the need for access revocation. As we have mentioned before, the amount of data in the storage system would increase as time goes by, and it would be rather time and energy consuming to do re-encryption. As it is hard for users to keep on alert for long times, the possibility of key exposition would increase day by day.

In order to relieve the user of remembering the details associated with storing their data, a centralized index might be maintained alongside the storage contents. All swords have two sides, besides the convenience it brought, the centralized index brings new threats to the system. As some systems are used by several users at the same time, a lot of information would be exposed to the attacker if he got a sufficient index of users. This places more long-term responsibilities on clients.

As the lifetime of media storage is extended to a long-term scale as decades, it gives attackers a much larger window to attempt to compromise the security system. With patience, an attacker can spend decades to implement an attack. It is very hard to detect slow attacks. One reason is that if the attack only makes the slightest changes at one time and each step is spaced apart far enough, it would be very difficult to detect the intrusion. Another reason is that it is hard to maintain attack history for a very long time, which is important for systems that use secret sharing algorithms [2]. For example, in secret sharing using a (3,5)-threshold scheme, a patient attacker can get one part per decade, and in less than three decades he would break the system and get the information he needs.

3.2 Threats to Integrity

Integrity means that only authorized entities can change the information or that the information can only be changed through an approved process.

In long-term secure storage, authentication must be considered with the time issues. Commonly, users need to show who they are before the system can decide what they can do, but in a long-term storage system, users primarily attached to the data may no longer be available. In that situation, a secure system must be able to authenticate new users and establish their relationships to the existing users and their resources.

In short lifetime storage, although data degrades as soon as it is placed on media, because the data is used frequently and updated regularly, the integrity of data in the system can be maintained. The integrity is checked in the form of disc auditing procedures, which periodically scans sections of data and uses strong hashes to detect corruption and to recover damaged ones. A variety of problems can arise by the auditing procedures: first, an overactive auditing procedure may lead to media failures; second, an under-active auditing procedure may not detect existing failures. As the opportunistic disk auditing policies requires auditing request to piggy-back on regular access requests, it is not sufficient for long-term storage.

On the other hand, cryptographic hashes may not be sufficient for integrity checking in the long term, as the attacker may find hash collisions used for disk auditing and update the hash data to fool the integrity checking procedure. Compare to the internal integrity problem of simple storage systems, there is an additional external integrity problem in distributed storage systems, which have to ensure that all the systems function properly. In such systems, one popular method of monitoring external integrity is using a challenge-response protocol, which could be easily employed by the attacker.

3.3 Threats to Availability

Availability means that information or a service is available to be used. In a long span of time, it is more possible to lose availability. There are a lot of threats to availability.

- First, large-scale disaster. During the lifetime of data, the system must take large-scale disasters into account, such as earth quakes and fire. Such disasters typically trigger other types of threat, such as media, hardware, and organizational faults.
- Second, human error. Users often accidentally delete content they still need, or purposefully delete data which they later discover a need. Sometimes the human errors may affect preservation of hardware, software, or infrastructure. Human error can increase system failures too.
- Third, component faults. Taking an end-to-end view of a system, any component may fail. Hardware components suffer transient recoverable and catastrophic irrecoverable faults. Software components, including firmware in disks, suffer from bugs affecting stored

data. Ingestion of data into a preservation system over the network may fail itself. External license servers or the companies that saved the data might no longer exist decades after.

- Fourth, media faults. The storage medium is a vital component. Almost all the affordable digital storage media are unreliable over long periods of time, and may be subject to gradual accumulation of irrecoverable bit errors or to sudden irrecoverable loss of bulk data.
- Fifth, hardware and software obsolescence. All hardware and software components can become obsolete or irreplaceable over time. This problem is particularly acute for removable media, which may be thrown away though readable. Software obsolescence is similar, the bits the data is encoded remain accessible, but the information can no longer be correctly interpreted.

4 Network File System

A network file system is any computer file system over a computer network, which supports sharing of files, printers and other resources as persistent storage. Normally, it is a client-server application which lets users view, store and update files on a remote computer as they are working on their own ones. Nowadays, network file systems become more and more popular in media storage. In this section we will discuss a number of network file systems and their security strategies.

4.1 pStore

pStore [5] is a secure distributed backup system based on an adaptive peer-to-peer network, which enables users to securely back up files into or get files from a distributed network of untrusted peers. It provides a large scale, massive storage system which provides scalability, high availability, persistence and security. In the pStore infrastructure, every node serves as access point for clients. Commonly, nodes are unreliable, as they may join the system and leave silently at any time.

Reliability is provided through replication of data. The copies of data are available on several different nodes in case some of them are malicious or unavailable. Security is ensured by using encryption and content hashes. As the data of the clients is replicated on nodes which out of their control, the security is assured as follows: private data is readable and can be deleted remotely by its owner only, and any unexpected changes can be detected easily.

When a file needs to be inserted by a user, an identifier which is specific and does not conflict with any other files or other users files is computed by pStore. After that the file is encrypted and broken into blocks that are digitally signed, and signed meta-data, which indicates how to reassemble the blocks, is assembled. Then both the meta-data and the blocks are inserted into the network.

4.2 OceanStore

OceanStore [1], a global, persistent distributed storage system, assumes that all of the nodes are untrusted and failure is inevitable. As it is very beneficial to disassociate the information from any fixed location and make it available anywhere, OceanStore is suitable for ubiquitous systems. In the system, only clients can be trusted with clear text. All information that enters the infrastructure must be encrypted.

OceanStore provides distributed access to persistent nomadic data in a uniform global scenario. As the servers are mainly untrusted in the system, data are protected through encryption and replication. This strategy provides high availability and it can prevent the system from denial-of-service attacks. OceanStore uses ACLs (Access Control Lists) to restrict write access to data, while read access is only available by using keys. High performance is provided by enforcing secrecy and integrity not only of the data in the system, but also of the meta-data and the lookup process. This can prevent the attack of the malwares.

4.3 The Coda File System

Coda [6], which has been developed at CMU (Carnegie Mellon University) since 1987, is a distributed filesystem with its origin in AFS2. Basically the security in Coda falls into two parts. The first one is Authentication and secure connections. The RPC2 package in Code can securely authenticate clients and servers, and set up encrypted channels between them. The password for users is the key element of this scheme. The second one is access control and protection database. The files on Coda servers are protect with directory access control lists, which grant permissions to users.

The control of access to files is simple, as the access control is established only after the establishment of an authenticated connection. The authentication and encryption process of data between clients and servers is really similar to the strategy used by Kerberos 5 (a network authentication protocol). However, it is not very qualified as it does not encrypt data on the server, this can be used by attackers as long as they break into the servers.

4.4 PAST

PAST [7] is a large scale P2P storage management system, which provides persistent storage with strong security by connecting peers on an overlay network. It achieves a high level of secrecy by using encryption on clients, and the authorization is accomplished through thoughtful use of certificates. Integrity and persistence characteristic is realized through the following approaches: as all data in the system is immutable, unwanted changes can be prevented. Randomized replication of data guarantees good availability.

The security model of PAST is based on the following assumptions: first, the public-key cryptosystem and the cryptographic hash function used in PAST are computationally infeasible to be broken; second, most nodes in the overlay network are well-functioning even if an individual PAST node is controlled by attackers; third, the smartcards hold by the users as a kind of certificate are out of the control of attackers.

The smartcards are used as a tool of certification, and they perform as trusted devices. The smartcard is used in many aspects of the read request process, which is based on fileIDs and routed by using the Pastry routing and location scheme[2]. The smartcard is also an essential token for system integrity from naming integrity to quota management. As the smartcard is so important, there should be a procedure for dealing with a lost smartcard, which is not published yet.

5 Conclusion

We have presented the archiving and security issues in trusted media storage above. Archiving, which includes the storage and retrieval of information, needs tight collaboration with annotations. In the media realm, annotation contains temporal, structured and free-form metadata, with emphasis on interaction techniques that improve the efficiency and accuracy of the annotation process. Security, which is a critical problem in all storage systems, becomes much harsher as a consequence of the long-term feature of media storage systems. As time itself becomes an active antagonist of storage, confidentiality, integrity and availability of the data may suffer more threats.

The requirements of trusted media storage are clear based on our own experience and that of others. As a common user, who does not need to know the details of the technologies being used, what they need most is the convenient management and great security.

Convenient management means to manage the data as fast as possible. The operation process should be as easy as possible. Nowadays, although a lot of archival systems have been proposed, seldom fulfill this requirement. As retrieval of data is managed through annotations, although there are some systems that can automatically attach metadata when the media is created, further works are still left for the users to do, because the auto-attached metadata are usually too simple to meet the users requirements. Due to the complexity of the automatic annotation technology, it is still not possible in the near future.

The great security indicates that the information should be secure no matter what happens. Obviously it is not possible. Nowadays, a lot of research has been done in the realm of secure long-term storage, although a lot of archival systems and network file systems have been proposed, barely of them specifically addressed all the security concerns. For example, the OceanStore, which is designed for long-term storage especially, still lacks strategy of dealing with hardware changes, migration of data and the slow compromises, not mention others.

In a word, there is still a lot of work left in the area of trusted media storage. In future it will be more sophisticated and successful.

References

- [1] John Kubiawicz, David Bindel, Yan Chen, OceanStore: An Architecture for Global-Scale Persistent Storage. Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000), November, 2000.
- [2] Mark W. Storer, Kevin Greenan, Ethan L. Miller, Long-Term Threats to Secure Archives. Conference on Computer and Communications Security, Proceedings of the second ACM workshop on Storage security and survivability, Alexandria, Virginia, USA SESSION: Studies and surveys, 2006, PP: 9-16.
- [3] Howard D. Wactlar and Michael G. Christel, Digital Video Archives: Managing Through Metadata. Environmental scan-Library of Congress' National Digital Information Infrastructure and Preservation Program. April, 2002.
- [4] Arnold W.M Smeulders, Franciska de Jong and Marcel Worring, Multimedia Information Technology and Annotation of Video. Audiovisueel: van emancipatie tot professionalisering. Stichting Archiefpublicaties, 's Gravenhage. Sponsored by MultimediaN DELOS, 2005, PP:95-115
- [5] Christopher Batten, Kenneth Barr, Arvind Saraf and Stanley Trepetin, pSore: A Secure Peer-to-Peer Backup System. Unpublished technical report MIT-LCS-TM-632. Massachusetts Institute of Technology Laboratory for Computer Science, Cambridge, MA. December, 2001
- [6] Peter Braam, Coda Authentication and Protection. Unpublished technical report of Carnegie Mellon University, Pittsburgh, PA <http://www.coda.cs.cmu.edu/doc/html/sec.html>
- [7] P. Druschel and A. Rowstron, PAST: A large-scale, persistent peer-to-peer storage utility. Eighth Workshop on Hot Topics in Operating Systems May, 2001, PP: 0075
- [8] Purdue University RASC: Confidentiality, Integrity and Availability (CIA) February 23, 2004 <http://www.itap.purdue.edu/security/files/documents/RASCCIAv13.pdf>

[1] John Kubiawicz, David Bindel, Yan Chen, OceanStore: An Architecture for Global-Scale Persistent Storage. Proceedings of the Ninth International Conference on Architectural Support for