A Study of the Use of IDAs in Cloud Storage

Xuesong Zhang and Honglei Wang

***Abstract*—Cloud storage is a model of networked online storage based on cloud computing, and provides users with immediate access to a broad range of resources and applications. Although a lot of cloud storage providers adopt encryption to protect costumer data, but users still suspect the security and privacy of their data. The paper analyzed Information dispersal algorithms (IDAs), and proved that it can better address the issues of confidentiality, integrity and availability of data. On this basis, the paper presented a cloud storage system adopting IDAs, and illustrated its key component and the process of writing file in cloud storage.**

***Index Terms*—Cloud storage, IDAs, security, proxy server.**

I. Introduction

Cloud storage is based on cloud computing, is the extension of Distributed Computing. Cloud storage is somehow the same as cloud computing, they consolidate all storage devices by the features like cluster application, grid technology and distributed file system[1].

At present, there are several wellknown cloud storage services. Amazon S3 (Simple Storage Service) is an online storage web service offered by Amazon Web Services. It provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web. It gives any developer access to the same highly scalable, reliable, secure, fast, inexpensive infrastructure that Amazon uses to run its own global network of web sites. The service aims to maximize benefits of scale and to pass those benefits on to developers[2]. Google Cloud Storage is a restful online storage web service for storing and accessing data on Google's infrastructure. It is an Infrastructure as a Service (IaaS), comparable to Amazon S3[3]. EMC Atmos is a cloud storage services platform developed by EMC Corporation. It can let enterprises and service providers store, manage, and protect globally distributed, unstructured content at scale. It provides the essential building blocks to implement a private, public, or hybrid cloud storage environment[4].

A. Cloud Storage Advantages

Cloud storage is a model of networked online storage where data is stored in virtualized pools of storage which are generally hosted by third parties. Hosting companies operate large data centers, and people who require their data to be hosted buy or lease storage capacity from them. The data center operators, in the background, virtualize the resources according to the requirements of the customer and expose them as storage pools, which the customers can themselves use to store files or data objects. Physically, the resource may span across multiple servers[10].

Cloud storage is made up of many distributed resources, but still acts as one. Cloud storage is highly fault tolerant through redundancy and distribution of data, and highly durable through the creation of versioned copies.

Companies need only pay for the storage they actually use as it is also possible for companies by utilizing actual virtual storage features like thin provisioning. They do not need to install physical storage devices in their own datacenter or offices, but the fact that storage has to be placed anywhere stays the same.

Cloud storage provides users with immediate access to a broad range of resources and applications hosted in the infrastructure of another organization via a web service interface[10].

B. Cloud Storage Problem

However, with third party managing the computing resources, off-site data storage raises several security and privacy concerns. Files distributed among a cluster of machines, often span national boundaries. It is not always possible to say under which jurisdiction and data protection laws the out-sourced information falls, and who will consequently be able to access it. Therefore, this usually means data owners lose the control over their data security and privacy. Companies feel nervous about their data, and do not sure whether their data has been adequately protected against theft from attackers both outside and inside the "cloud". Particularly sensitive customer and personal data need to be protected and are subject to heavy constraints that cannot always be matched by current cloud storage solutions. Although a lot of cloud storage providers adopt the most convenient way, encrypting data, in order to protect costumer data, but they usually do the key management themselves. As a consequence, users have no influence on the file encryption process and lose control over who may have access to their data [5]. To ensure confidentiality of data, Customers usually have to select a most reliable cloud storage supplier, or resort to encrypting their data before sending them to cloud storage device. To ensure availability and integrity of data, Customers normally replicate their data. However these measures have their limitations in a public cloud environment.

Storage system which is based on information dispersal will better address the issues of confidentiality, integrity and availability of data [10]. Current research and commercial implementations storage systems employing IDAs are more suited for static file storage and retrieval. CleverSafe[12], focuses on storage, using information dispersal to slice TCP/IP packets and store them on a network of local or

Manuscript received April 23, 2012; revised May 25, 2012.

The authors are with Xuzhou College of Industrial Technology, Jiangsu, China (e-mail: zxscedar1@ 163.com, 98312226@163.com).

remote servers in a system called Dispersed Storage. The company's Accesser product slices the data using IDAs [10].

"Information dispersal algorithms (IDAs) have the ability to disperse data in a very secure way across a number of nodes so that if you compromise one node, you won't compromise any data," said Michael Versace, a Wikibon Project partner and analyst. "We're hearing there are a lot of people looking at IDAs as a replacement or an alternative to traditional data encryption."[10]

II. Overview of Information Dispersal Algorithms

Information dispersal algorithms (IDAs) – first proposed by algorithm researcher Michael O. Rabin in 1989 – are used to slice data into pieces at the bit level so that when data traverses the network or sits in storage arrays, it is unrecognizable unless accessed by a user/device with the right key. When accessed with the right key, the information is reassembled. [10]

A. The Main Algorithm

In IDA, a file is split into n slices and a minimum of m slices $(n > m)$ are required for reconstructing the original file. A transform matrix of n rows and m columns is used to perform the transformation.[6]

Let F be the original file of size N and be a byte array. The bytes in F can be chunked into blocks of m bytes, as in (1):

$$F = (b_1, b_2, b_3, \ldots b_m), (b_{m+1}, b_{m+2}, \ldots b_{2m}) \ldots (b_{N-m+1}, \ldots b_N) \quad (1)$$

Let A be the transform matrix, as in (2):

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{bmatrix} \quad (2)$$

Let B be the input file matrix of F and C the output file matrix. Therefore, the following matrix equation can be obtained, as in (3):

$$A \bullet \begin{bmatrix} b_1 & b_{m+1} & \ldots & b_{N-m+1} \\ b_2 & b_{m+2} & \ldots & b_{N-m+2} \\ \vdots & \vdots & \ddots & \vdots \\ b_m & b_{2m} & \ldots & b_N \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1N/m} \\ c_{21} & c_{22} & \ldots & c_{2N/m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \ldots & c_{nN/m} \end{bmatrix} \quad (3)$$

To get c_{11}, because c_{11} equals to row 1 of A multiplied by column 1 of B, the following equation can be used, as in (4):

$$c_{11} = a_{11}b_1 + a_{12}b_2 + \ldots + a_{1m}b_m \in GF(2^8) \quad (4)$$

Each row of C corresponds to a slice.

Assume the slices 1 to m to be used for recombination. Let A' be a subset of A of rows 1 to m. Let A^{-1} be the inverse matrix of A', as in (5):

$$A^{-1} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mm} \end{bmatrix} \quad (5)$$

For reconstruction, the inverse matrix A^{-1} is applied to the m (out of n) slices to obtain the original data, as in (6):

$$A^{-1} \bullet \begin{bmatrix} c_{11} & c_{12} & \ldots & c_{1N/m} \\ c_{21} & c_{22} & \ldots & c_{2N/m} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \ldots & c_{mN/m} \end{bmatrix} = \begin{bmatrix} b_1 & b_{m+1} & \ldots & b_{N-m+1} \\ b_2 & b_{m+2} & \ldots & b_{N-m+2} \\ \vdots & \vdots & \ddots & \vdots \\ b_m & b_{2m} & \ldots & b_N \end{bmatrix} \quad (6)$$

B. Security in Cloud Storage

In Cloud Storage, IDAs can be employed to split files into multiple data slices which will be redundantly stored on several storage nodes. IDAs can also enhance the stored data confidentiality. In a cloud storage system employing IDAs, a adversary who wants to read a file, has to compromise minimum slice stores or eavesdrops on slices [3]. He/she also needs to determine which slices logically belong to the file. The adversary will also have to guess the transform matrix of the file and apply the matrix with the correct sequence of the slices. To achieve all these will be very difficult, in practical. The probability of success is very little. Therefore, the storage system employing IDAs, rather than that just employing encryption, can more effectively guarantee the confidentiality of the data.

To further enhance the security of IDAs-based Cloud Storage, the proxy server can optionally encrypt the file slices before sending them to external storage services. The system can be configured to use a randomly generated matrix for every file, instead of using a global matrix for all the files [6].

III. Cloud Storage System Using IDAs

To overcome the security and privacy issues with storing data in cloud computing network as they were discussed in previous sections, we propose a system architecture adopting IDAs for securing off-site data storage, depicted in Figure 1.

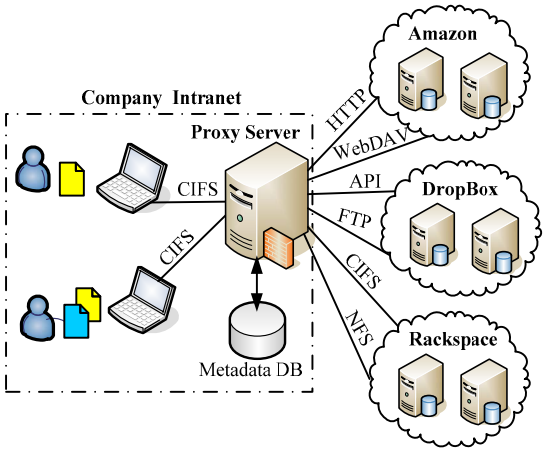


Fig. 1. System architecture.

A. Proxy Server

The key component of this system is a proxy server which is responsible for integrating the external storage services from the Internet, offering the new resources to the client computers on the intranet, and securing all data transfers as soon as they leave the trusted enterprise network zone[7].

The proposed proxy server can be a Linux based system , and usually be located within the trusted zone of a company's intranet. One of the major goals is to seamlessly integrate the

external cloud storage services into an employee's desktop work-space using the proxy server as a mediator. Therefore, the user no longer has to add any additional software components for storing files in Cloud Storage. Through Common Internet File System(CIFS), the user can employ the additional storage resources offered by the proxy server on the client computer. CIFS is a protocol that lets programs make requests for files and services on remote computers on the Internet [1]. It uses the client/server programming model , and seems to be the best choice in the mostly Windows dominated world of office computers [5].

For the purpose of starting the dispersion and encryption algorithms on the proxy server, a customized file system is necessary that enables users to "overwrite" the standard file system operations. The famous Filesystem in Userspace(FUSE) will be adopted here to achieve the necessary functionality [11]. FUSE is a loadable kernel module for Unix-like computer operating systems that lets non-privileged users create their own file systems without editing kernel code. [9].

B. Write Operations

The process of writing file in Cloud Storage will be carried out as follows, depicted in Figure 2: the user usually a company employee copies a file to a desired folder on the network drive. This file will be cached on the proxy server, almost at the same time the proxy server will generate a special matrix for the file randomly. Then the proxy server will use the matrix to transform the file into multiple slices. In order to further enhance the confidentiality of information, the data fragments will be encrypted by the proxy server before they leave the trusted intranet. At last, the resulting data slices will be stored on the online cloud drives provided by Amazon, Dropbox, or Rackspace via the protocol adapter [5].

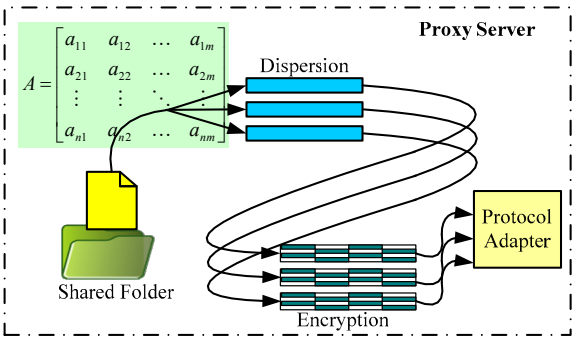


Fig. 2. Writing process.

C. Metadata DB

In order to maintain the mapping of file to file slices, class FileInfo and class FileSlice are used as shown in Figure 3. Class FileInfo contains properties such as IDA matrix used to slice the file, namespaces of slice stores, as well as general file attributes and so on [5]. Class FileSlice contains information of one file slice. A file can be divided into a number of file slices. All file instances are stored in metadata database. During the whole writing process additional information and metadata belonging to the out-sourced file will be stored into the database which allows the cached file to be deleted from the proxy server after the storage procedure was completed successfully [6].

By employing metadata database, proxy server can manage slice store registration and removal, as well as keeping track of the health and availability of all the slice stores [6]. The database is going to be a MySQL db, which should be replicated, distributed, and protected against attacks and failures according to best practices in the field of database security [5].

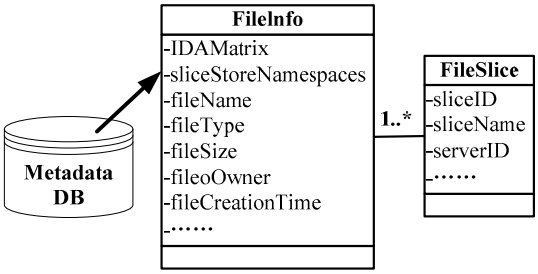


Fig. 3. Metadata DB.

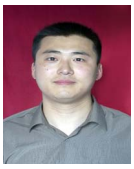
IV. Conclusion

In this paper, we presented a system architecture that provides data security, integrity and availability suited for deployment in a public cloud environment. The system's core component is a proxy server which is responsible for data distribution via IDAs. In order to further enhance the data security, the proxy server will generate a special matrix for each file randomly, and will encrypt file slices before sending them to cloud storage services. Due to this fact that all operations are performed in the trusted local intranet, the usual security concerns and privacy issues with common cloud storage services should be mitigated.

References

[1] J. Wu, J. Fu, L. Ping, and Q. Xie, "Study on the P2 P Cloud Storage System," *ACTA* Electronica *Sinica*, vol. 139, no.5, pp. 1100-1106, May 2011.
[2] Amazon Simple Storage Service (Amazon S3). [Online]. Available: http://aws.amazon.com/s3/.
[3] What is Google Cloud Storage? [Online]. Available: https://developers .google.com/storage/.
[4] EMC at a Glance. [Online]. Available: http://www.emc.com/storage/atmos/atmos.htm.
[5] R. Seiger, S. GroB, and A. Schill, "SecCSIE: A Secure Cloud Storage Integrator for Enterprises," in *Proc. IEEE Conf. on Commerce and Enterprise Computing*, 2011, pp. 252-255.
[6] K. K. Mar, "Secured virtual Diffused File System for the Cloud," in *Proc. 6th Int. Conf. on Internet Technology and Secured* Transactions, United Arab Emirates, 2011, pp. 116-121.
[7] S. Luo, F. Liuz, and C. Ren, "A hierarchy attribute-based access control model for cloud storage," in *Proc. Int. Conf. on Machine Learninh and* Cybernetics, China , 2011, pp. 1146-1150.
[8] L. Xu, J. Hu, S. Mkandawire, and H. Jiang, "SHHC: A scalable hybrid hash cluster for cloud backup services in data centers," in *Proc. 31st Int. Conf. on Distributed Computing* Systems *Workshops*, 2011, pp. 61-65.
[9] H. Qinghua, W. Yongwei, Z. Weimin, and Y. Guangwen, "A Method on Protection of User Data Privacy in Cloud Storage Platform," *Journal of Computer Research and Development*, vol. 48, no. 7, pp.1146-1154, May 2011.
[10] Information dispersal algorithms: Data-parsing for network security. [Online]. Available: https://developers.google.com/storage/.
[11] Filesystem in Userspace. [Online]. Available: http://en.wikipedia.org /wiki/Filesystem_in_Userspace.
[12] Cleversafe and QStar Technologies Form Technology Partnership. [Online]. Available: http://www.prnewswire.com/news-releases
[13] /cleversafe-and-qstar-technologies-form-technology-partnership-1475 65325.html.

Xuesong Zhang received the first bachelor's degree in chemical equipment and machine from Nanjing University of Technology in 2001. She received the second bachelor's degree and master's degree in computer science and technology from China University of Mining & Technology in 2003 and 2008, respectively. She is currently a lecturer in Xuzhou College of Industrial Technology in China. Her research interests include software engineering, database applications, and the application of computer and operations research in engineering



Honglei Wang received the bachelor's degree in electronic engineering from Jiangsu Teachers University of Technology in 2002, and received the master's degree in communication and information system from China University of Mining & Technology in 2009. He is currently a lecturer in Xuzhou College of Industrial Technology in China. His research interests include data communication, wireless communication, embedded system design, and the cloud computing networks research in application.