ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

An Investigation on Image and Data Storage in Cloud Environment with an Enhanced Approach of Data Compression using Compressed Sensing

Manju Sadasivan [1], Radha Sridharan [2], Ranjitha M[3]

Asst. Professor, Dept of Information Technology, CMR Institute of Management Studies, Bangalore, India[1, 2, 3]

Abstract: High definition images are generated using the latest technology in image capturing. However the storage space requirements increases manifold since the large size of high definition images demands more memory. There have been a few methodologies recommended by various studies for the storage and retrieval of image data. This paper has identified two popular approaches to store images on cloud infrastructure. The first approach is to make use of Mirage Library which has commonly been exercised in IaaS Clouds to store image as structured data. The next approach is based on a technology called Deduplication to store the image by deploying kernel-space file system with Deduplication in the image storage server. Both these approaches are found to be effective in reducing the storage needs on virtual image store. This paper discusses a technique to compress data chunks (that are pre-processed through Mirage Library or Deduplication) using Compressed Sensing to further reduce the storage needs. It also encrypts the data chunks to ensure secure data storage.

Keywords: Deduplication, Mirage, Delta deployment, Image storage, Open Stack, Compressed Sensing.

I. INTRODUCTION

Cloud Computing enables universal, expedient network access to a shared area of configurable computing resources like networks, servers, storage, applications, and services that can be quickly managed with least effort [1]. It is a computational model as well as a distribution style and its main aim is to provide secure, quick, convenient data storage and net computing service. It also facilitates the growing demand for information storage by emphasizing on cost reduction [2]. The past decade has seen tremendous change and volatile growth in the management of data which has resulted in a huge demand for a technique to store these data in an efficient way. Cloud Computing is a widely accepted technology to solve this problem. It offers a wide variety of techniques to deal with unstructured data which contains images as well [3].

A virtual machine image gives a virtual representation of a client's machine and thereby cloud computing systems generate large number of Virtual Machine (VM) Images. In addition to the storage capacity, image transmission overhead such as transmission time and storage time has to be given due importance. Therefore managing these massive images will be the future challenge. File systems are designed into two major categories- designed for internet services, application that executes in parallel on large configurations [4]. This work concentrates on two major approaches of managing images and data i.e. by using mirage library and deduplication which is discussed in the following sections.

II. IMAGE STORAGE AND RETRIEVAL USING MIRAGE LIBRARY

Glenn Ammons et al [5] proposes a methodology for image storage and retrieval in the form of Mirage Library which is one of the typically used image library in IaaS clouds that stores images as structured data. A virtual machine image gives a virtual representation of a client's machine and thereby cloud computing systems generate large number of Virtual Machine (VM) Images. The problems in managing large collections of these VM images in IaaS clouds are addressed by Mirage Library. One of the unique features of this library is that it allows introspection and manipulation of images in an offline environment. It converts the image into a file aware format using indexed file system structure that exposes the internal structure of the image. This feature enables easy governance, maintenance, searching and comparison activities. This is accomplished using check-in and check-out interfaces that convert disk images into a structured file format in and out of mirage library. The components of mirage library are:

A. Content Addressed Store

This allows storage of arbitrary data. Each item in the store is identified uniquely by an identifier. Identical items are referred by the same identifier to avoid duplication of storage as the items are stored only once. The semantics of CAS (Content Addressed Store) is writing-once which implies that any modified item will have a different identifier. The structure of the image is stored as a hierarchical tree structure where the leaf node stores the contents of the file, the internal nodes contain the disk manifest for disk images and image manifest for VM images. Once the identifier of the root of the structure (an image manifest) is known, it can securely identify the entire structure [6].

B. Image Indexer

The image indexer converts disk images to mirage format and vice versa. By using hybrid indexing, it supports different types of disk formats. While converting a disk

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

image to mirage format and back, it does not produce a bit-by-bit identical copy of the original. Instead, it restores the image by creating empty file systems which can appear in different disk blocks. It also reconfigures the affected boot loaders at check-out time.

C. Catalog manager

The catalog manager stores meta-data (data about the data) about images in a RDBMS format that includes CAS identifier, time of creation, whether it is derived from other image, the parent's identifier etc. It also maintains version information in the form of named images and provides controlled access to these named images. A typical Mirage Architecture is shown in Fig 1.

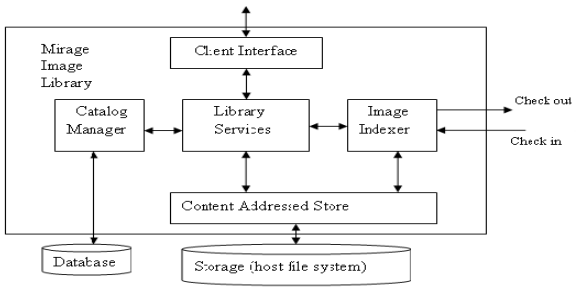

Fig 1: Mirage Architecture [5]

D. Library services

The library services component provides the user-visible and administrator-visible functions of Mirage. The user-visible functions include checkin, checkout, version control, virtual mount and analytics functions (describing, listing, comparing, and searching images). The administrator-visible functions include starting or stopping the Mirage server, controlling garbage collection, locking selected images for maintenance, and other administrative tasks.

E. Mitigating translation costs

The translation cost that is imposed while converting disk images to file-aware mirage format is greatly reduced by factors like:-

1) A structure-aware CAS: ensures arbitrary storage of data.

2) Virtual mount: read; write operations on images as patches can be performed offline by mounting an image's file system without reconstructing the images disks.

3) Delta deployment: makes use of similarity between images and hence reduces disk recreation costs. Delta deployment calculates file-level deltas between pairs of similar disk images. This file-level delta specifies the files that should be added, modified or deleted in the source image so that target image can be produced without recreating them. This is implemented by keeping a cache of popular disk images. When an image is retrieved and if it is present in the cache, a clone is created instead of recreating the disk image. If the target image is not found in the cache, delta deployment searches for the source image. If it exists, a clone is created, the calculated file-level delta is applied to the clone so that the target image is created.

4) Hybrid indexing: Mirage supports variety of file systems. It uses indexing of image disks in the form of fixed size disk blocks.

III. IMAGE STORAGE USING DEDUPLICATION TECHNOLOGY

Jian Wan et al [7] proposes a methodology for image management in cloud using Deduplication technology in open stack. Tremendous image data bring greater pressure on storage space and network transmission time. Increase in storage efficiency, decrease in hardware costs, cost reduction for backups and the disaster recovery costs reduction are the advantages of Deduplication.

Deduplication is used to control data commonality in a storage system. Controlling is done by identifying duplicate "chunks" of data across multiple files and storing only one copy of each chunk. By using the above method 80% usage of the "Virtual Machine Image Storage" is reduced. This method applies static block technology to divide file into a number of blocks of data. File chunking, fingerprint extraction, fingerprint lookup and data storage are the four steps in Deduplication process. Fixed size chunking and Variable size chunking are the two methods to break a file into sections or blocks.

This paper focus on the "Fixed chunking technique". After dividing the file into fixed size blocks, next step is to calculate the fingerprint using the fingerprint program. The calculated fingerprint is then pushed to the fingerprint filter to look for identical fingerprint, implying redundant data block. Duplicated blocks are replaced with reference to the stored block. The system then writes the block into the disk and modifies the inode. Compared to conventional compression tools, this method yields better results.

The flow of operation of the Deduplication system is depicted in Fig 2.

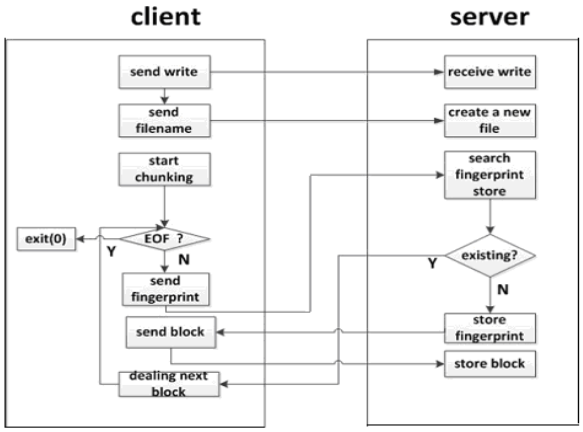

Fig 2: Flow of operations of the system [7][8]

A. Technologies used

1) Open stack: The key components of Open stack includes - nova, glance, keystone, swift, network etc. Nova performs computation while Glance provides image management functionality. Keystone provides registration and management for the account [7][8]. Open stack does two things when storing an image: (i) storing a virtual image in the glance (ii) registering the image information in the database.

2) Image De-duplication: The system combines the features of liveDFS and uses fixed size chunking. This data de-duplication method is used in the image data storage on a single node.

DOI 10.17148/IJARCCE.2015.41056

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

3) Fingerprint Filter: The fingerprint filter is used to speed up the finding speed of the fingerprint on disk, which is the index structure in memory. It is a two level procedure. The first-level filter map the former 'n' bit of fingerprint which is called as index key. The second-level filter maps the later k bits which are called as bucket key. The finger print filter structure is shown in Fig 3.

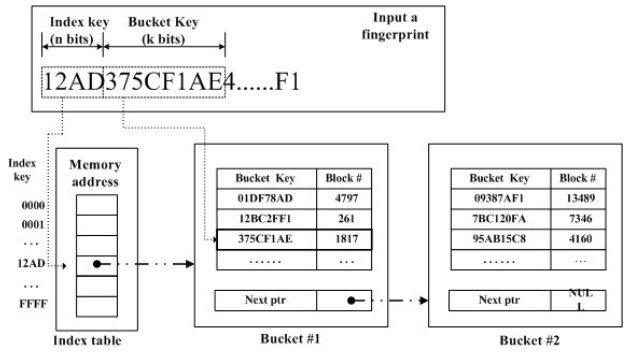

Fig 3: The structure of Fingerprint Filter [5]

The index key table is an array containing 2n elements. Each element in this array points to the header of a bucket linked list. The tuples in the bucket are pointed by this index key. Tuples key block number is combined with the bucket line. Once bucket is full, a new bucket linked list is created. The elements in each bucket are sorted to query the bucket key [7].

4) Fingerprint Calculation: After dividing the file into fixed size chunks, each file chunk need to calculate the data fingerprint. System then sends the fingerprint and other meta data to the image storage management server [7]. On the server side, the fingerprint is retrieved to identify that whether it is required to send the data to image storage server. The two algorithms which are used to generate fingerprint are: Bloom Filtering Algorithm and Hashing Algorithm (MD5, SHA1).

B. System Design

In Open stack, nova integrates physical resources from compute nodes and provides users computing services. Glance manages the registration, storage and retrieval of virtual machine image. The glance server will initially record the image information in the registry server and uploads the image to the storage backend which is demonstrated in Fig 4.

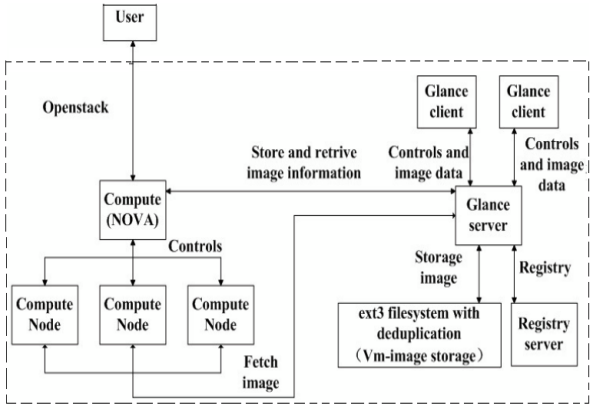

Fig 4 Architecture Diagram [7]

To implement this architecture three servers are used: Original image storage server, Deduplication image storage server and a client machine. Deduplication rate depends primarily on Deduplication algorithm and Deduplication data sets.

IV. SECURITY ISSUES

Darapaneni Chandra Sekhar et al [9] has proposed a technique that addresses the security issues while storing and retrieving image on cloud environments. High definition images which provide vital information like medical images, remote sensing images, satellite image databases, etc. are outsourced to cloud so that images can be shared between data owners and large sets of data users.

To protect the privacy-sensitive data, a compressed sensing technology is adopted. Compressed image samples can be easily captured on the storage servers by data owners via a simple non-adaptive linear measurement process from physical imaging devices, and later easily be shared with users. Compressed sensing also helps in storage overhead reduction because the size of the compressed sample is always less than the size of the actual image. The storage cost is reduced by 50% because of compressed sensing.

TISR (Trusted Image Storing and Retrieval Framework)is specifically designed under the compressed sensing framework to establish secure and privacy-assured image service outsourcing in cloud computing. Fig 5 shows the architecture of TISR.

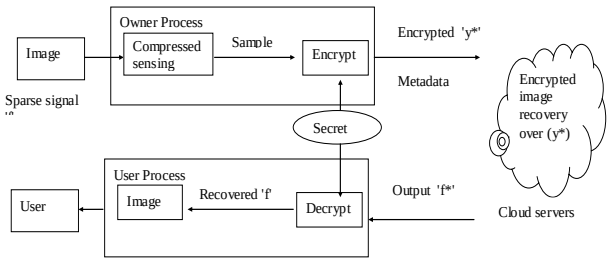

Fig 5: Architecture of TISR [6]

First raw image data as compressed image samples are captured by the data owner from physical imaging devices. These compressed raw image samples are published in cloud for storage and processing. Whenever there is a demand for these images, the cloud reconstructs them. Since these images are in compressed form, it requires lot of computation to reconstruct the original image. Moreover while reconstructing the images the cloud may be curious to know about the owner's/user's data .Since these are privacy sensitive data, TISR ensures that data is protected once it moves out of owner's/user's process.

TISR handles this challenge through random transformation based framework that includes

1) PGeneration : algorithm running at the data owner end, for generating the secret key

2) PTrans : algorithm running at either data owner or data user that takes the secret key P as input and generates a randomly transformed problem Ω p where Ω is the original problem.

DOI 10.17148/IJARCCE.2015.41056

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

3) PSolv : algorithm running at cloud side, solves the problem Ω p and generates 'h'

4) PRec : algorithm running at the data user gets P and h and generates the answer g of the original problem

We denote this framework of TISR as r = (PGen, PTran, PSolv, PRec).

V. RESULTS AND DISCUSSIONS

In our comparative approach, we have studied on these three methods and the observations are:

Glenn Ammons's technique of storing images in Mirage Library ensured that disk images are represented in cloud as structured data in indexed file systems. The other benefits of this approach are lower translation costs from disk images to VM images and vice versa, delta deployment where file level deltas were used to identify only those portions of the images that were modified were sent as patches to cloud instead of reconstructing the whole image.

Jian Wan's methodology of storage in cloud implements Deduplication to reduce storage memory. It uses memory filter to reduce the number of disk index. It centralizes fingerprint storage area to improve the locality of data accessing and uses limited memory to achieve a higher IO throughput rate [7].

Darapaneni Chandra Sekhar's approach suggests a trusted, secured methodology of storing images through Compressed Sensing. By hiding the underlying image content through compression, data is secured while storage and retrieval at both the cloud end and data user end.

VI. CONCLUSION

This work has analyzed the two major approaches used for image storage and retrieval in cloud. We have observed that the first two approaches can still save their storage space and memory by incorporating 'Compress sensing' technique into their methodology. By incorporating this technique, the security issues are addressed to some extent.

REFERENCES

[1] http://www.nist.gov/itl/cloud/
[2] http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_12-/123_cloud1.html.
[3] James Manyika Michael Chui Brad Brown Jacques Bughin Richard Dobbs Charles Roxburgh Angela Hung Byers," Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011.
[4] Bin Cui1, Hong Mei, Beng Chin Ooi2,"Big data: the driver for innovation in databases" ,National Science Review, Volume 1, Issue 1, pp. 27-30,2013.
[5] Glenn Ammons, Vasanth Bala, Todd Mummert, Darrell Reimer and Xiaolan Zhang, "Virtual machine images as structured data: the Mirage image library", IBM Research.
[6] Quinlan, S., Dorward, S. Venti, "A new approach to archival storage", Proceedings of the FAST '02 Conference on File and Storage Technologies, pp. 89–101, 2002.
[7] Jian Wan, Shuting Han, Jilin Zhang, Baojin Zhu and Li Zhou ," An Image Management System Implemented On Open-source Cloud Platform", IEEE Proceedings of International Symposium on Parallel & Distributed Processing Workshops and PhD Forum, pp.2065 – 2070, 2013.
[8] Jilin Zhang, Shuting Han, Jian Wan, Baojin Zhu, Li Zhou, Yongjian Ren, and Wei Zhang, "IM-Dedup: An Image Management System Based on Deduplication Applied in DWSNs", International Journal of Distributed Sensor Networks, Volume 2013, 2013
[9] Darapaneni Chandra Sekhar , M.Chandra Naik – "An Enhanced trusted Image Storing and Retrieval Framework in Cloud Data Storage Service Environment", IJERA , Volume 4, Issue 10 , pp . 84 -88, 2014.