INDEX BASED DEDUPLICATION FILE SYSTEM USING DISTRIBUTED HASH TABLE

Prabakaran U
Computer Science and Engineering
Vandayar College of Engineering and Technology
Thanjavur, India

Subbulakshmi G
Head of the Department
Computer Science and Engineering
Vandayar College of Engineering and Technology
Thanjavur, India

ABSTRACT:

In a virtualized cloud environment each instance of a guest operating system runs on a virtual machine, contact virtual hard disks symbolized as virtual disk image files in the host operating system. Because these image files are stored as regular files from the external point of view, backing up VM's data is mainly done by taking snapshots of virtual disk images. In a virtualized cloud computing surroundings, common snapshot support of virtual disks improves hosting reliability but storage demand of such operations is huge. While dirty bit based technique can identify unmodified data among versions; full deduplication with fingerprint assessment can remove more redundant content at the cost of computing resources. So in this project we proposed a LIQUID framework to implement the deduplication scheme in both file and block level. And propose a deduplication file system with low storage consumption and high-performance IO, which satisfies the requirements of VM hosting. Finally we extend our work to address the problem of authorized data deduplication.

Index terms: ***Deduplication, cloud computing, virtual machine, liquid, file system***

1. INTRODUCTION:

Cloud computing offer seemingly limitless "virtualized" income to users as services across the complete Internet, while thrashing platform and implementation details. Todays cloud service supplier offer both highly accessible storage and particularly parallel computing resources at comparatively small costs. As cloud computing becomes common, a rising amount of data is being store in the cloud and shared by users with particular rights, which describe the way in rights of the stored data. For years the Internet has been symbolized on network diagrams by a cloud symbol until 2008 when a diversity of new services in progress to appear that allowable computing resources to be right of entry over the Internet term Cloud Computing.

Cloud computing include actions such as the use of social networking sites and other forms of interpersonal computing; though, most of the time cloud computing is disturbed with right to use online software applications, data storage and processing power. Cloud computing is a way to enlarge the capacity or add capability dynamically with no spend in new infrastructure, training new personnel, or licensing new software. It expands Information Technologies (IT) obtainable capability. In the last few years, cloud computing has developed from being a talented business idea to one of the fast rising segments of the IT industry. Cloud computing is an on-demand Internet-based virtual service model where data is stored and access utilize distributed on-demand elastic services in a transparent way. Internet has developed by bounds and bounds and is rising at an inconceivable pace. Each business on the planet has its personal on the Internet. With the arrival of cloud computing, business houses have establish it simple to host their applications on a virtual situation hosted and directed by a cloud service supplier. The idea of on-premise application use is fast vanishing out. This is recognized as Software-As-A-Service (SaaS) in cloud computing world. This gives organizations the flexibility to decide the best cloud hosting supplier support on parameters such as cost, dependability, presentation, storage, security, accessibility etc. One serious dispute of cloud storage services is the organization of the ever- increasing volume of data. To create data management scalable in cloud computing, deduplication has been a well-known method and has involved more and more concentration lately. Data deduplication is a particular data compression method for removing duplicate copies of replicate data in storage. The method is utilized to get better storage use and can also be applied to network data transfers to decrease the number of bytes that must be sent. As an alternative of keeping numerous data copies with the similar contented, deduplication reduces redundant data by maintaining only one physical copy and referring other redundant data to that copy. Deduplication can obtain place at either the file level or the block level. For file level deduplication, it removes duplicate copies of the similar file. Deduplication can also obtain place at the block level, which removes duplicate blocks of data that happen in non-identical files. Although data deduplication carry a lot of benefits, security and privacy apprehension arise as users□ responsive data are vulnerable to both inside and outside attacks. Traditional encryption, while provided that data privacy is unsuited with data deduplication. Particularly, traditional encryption needs dissimilar users to encrypt their data with their own keys. Thus, identical data copies of dissimilar users will guide to dissimilar cipher texts, creating deduplication not possible.

2. RELATED WORK:

In [1] A. Mathur, M. Cao, S. Bhattacharya, A. Dilger, A. Tomas, and L. Vivier. Ext3 has been the most widely used general Linux RO file system for many years. In keeping with increasing disk capacities and state-of-the-art feature requirements, the next generation of the ext3 file system, ext4, was created last year. This new file system incorporates scalability and performance enhancements for supporting large file systems, while maintaining reliability and stability. Ext4 will be suitable for a larger variety of workloads and is expected to replace ext3 as the "Linux file system." In this paper we will first discuss the reasons for starting the ext4 filesystem, then explore the enhanced capabilities currently available and planned for ext4, discuss methods for migrating between ext3 and ext4, and finally compare ext4 and other file system performance on three classic file system benchmarks.

In [3] C. Ng, M. Ma, T. Wong, P. Lee, and J. Lui. Deduplication is an approach of avoiding storing data blocks with identical content, and has been shown to effectively reduce the disk space for storing multi-gigabyte virtual machine (VM) images. However, it remains challenging to deploy deduplication in a real system, such as a cloud platform, where VM images are regularly inserted and retrieved. We propose LiveDFS, a live deduplication file system that enables deduplication storage of VM images in an open-source cloud that is deployed under low-cost commodity hardware settings with limited memory footprints. LiveDFS has several distinct features, including spatial locality, prefetching of metadata, and journaling. LiveDFS is POSIX-compliant and is implemented as a Linux kernel-space file system. We deploy our LiveDFS prototype as a storage layer in a cloud platform based on OpenStack, and conduct extensive experiments. Compared to an ordinary file system without deduplication, we show that LiveDFS can save at least 40% of space for storing VM images, while achieving reasonable performance in importing and retrieving VM images. Our work justifies the feasibility of deploying LiveDFS in an open-source cloud.

In[9] Stoica, R.Morris, D. Karger,M.F. Kaashoek, and H Balakrishnan. A fundamental problem that confronts peer-to-peer applications is to efficiently locate the node that stores a particular data item. This paper presents Chord, a distributed lookup protocol that addresses this problem. Chord provides support for just one operation: given a key, it maps the key onto a node. Data location can be easily implemented on top of Chord by associating a key with each data item, and storing the key/data item pair at the node to which the key maps. Chord adapts efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing. Results from theoretical analysis, simulations, and experiments show that Chord is scalable, with communication cost and the state maintained by each node scaling logarithmically with the number of Chord nodes.

3. DEDUPLICATION SYSTEM:

For VM snapshot support, file level semantics are usually not offered. Snapshot operations obtain position at the virtual device driver level, which means no fine-grained file system metadata can be utilized to establish the distorted data. Backup systems have been urbanized to utilize contented fingerprints to recognize duplicate contented. Offline deduplication is utilized to eliminate before written duplicate blocks through inactive time. Several methods have been proposed to speedup seeking of duplicate fingerprints. Existing methods have alert on such inline duplicate discovery in which deduplication of an entity block is on the significant write path. In existing work, this restraint is complicated and there is no waiting time for a lot of duplicate discovery requests. This recreation is offensive because in circumstance, complicated to finishing the backup of necessary VM images within a sensible time window.

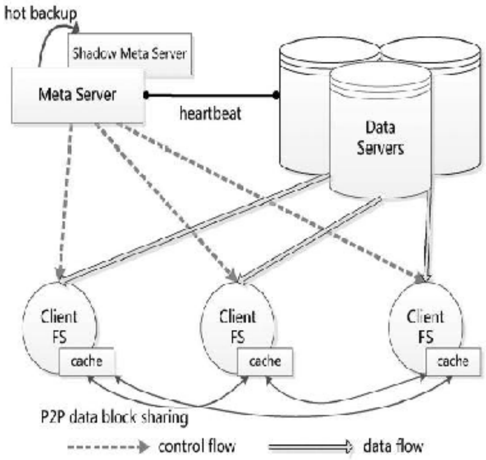

Figure 1: liquid architecture

4. TRUST BASED P2P DATA SHARING DEDUPLICATION IN CLOUD COMPUTING

In paper proposed centralized backup structural design with multi-level and selective deduplication. This service is hosted in the existing set of machines and resource procedure is controlled with a negligible collision to the existing applications. The deduplication procedure is first execute between snapshots inside every VM and then is perform across VMs. Given the apprehension that seeking duplicates across VMs is a global feature which can involve similar presentation and confuse failure management, only abolish the duplication of a small but accepted data set while still preserve a cost-effective deduplication ratio. For this reason, use the data characteristics of snapshots and gather most accepted data. Data sharing across VMs is limited inside this small data set such that addition replicas for it could improve fault tolerance.

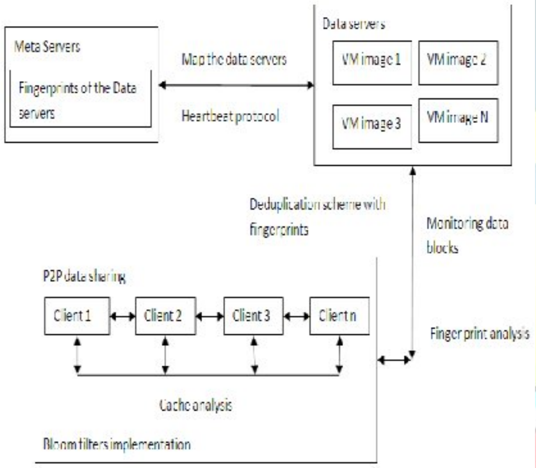

Figure 2: System architecture for proposed

5. CONCLUSION:

In cloud many data are stored again and again by user. So the user need more spaces store another data. That will reduce the memory space of the cloud for the users. To overcome this problem uses the deduplication concept. Data deduplication is a technique for reducing the amount of storage space an organization wants to save its data. In many associations, the storage systems surround duplicate copies of many sections of data. For instance, the similar file might be keep in several dissimilar places by dissimilar users, two or extra files that aren't the same may still include much of the similar data. Deduplication remove these additional copies by saving just one copy of the data and replacing the other copies with pointers that lead back to the unique copy. Companies regularly utilize deduplication in backup and failure improvement applications, but it can be utilized to liberate up space in main storage as well. In its simplest form, deduplication takes place on the file level; that is, it removes duplicate copy of the similar file. This type of deduplication is occasionally called file- level deduplication or single instance storage (SIS). Deduplication can also obtain place on the block level, removing duplicated block of data that happen in non- identical files. Block-level deduplication frees up more spaces and a exacting category recognized as variable block or variable length deduplication has become very popular. Often the phrase "data deduplication" is used as a synonym for block-level or variable length deduplication. In cloud using the DHT table the user easily searches the data and retrieves the searched data from the cloud.

6. REFERENCE:

[1] A. Mathur, M. Cao, S. Bhattacharya, A. Dilger, A. Tomas, and L. Vivier, „„„The New ext4 Filesystem: CURRENT STATUS and Future Plans,□□ in Proc. Linux Symp., 2007, vol. 2, pp. 21-33, Citeseer.

[2] M. McLoughlin, The qcow2 Image Format, Sept. 2011. [Online]. Available: http://people.gnome.org/markmc/qcow-image-format.html

[3] C. Ng, M. Ma, T. Wong, P. Lee, and J. Lui, „„„Live Deduplication Storage of Virtual Machine Images in an Open-Source Cloud,□□ in Proc. Middleware, 2011, pp. 81- 100.

[4] K. Pepple, Deploying OpenStack. Sebastopol, CA, USA: O□Reilly Media, 2011.

[5] S. Quinlan and S. Dorward, „„„Venti: A New Approach to Archival Storage,□□ in Proc. FAST Conf. File Storage Technol., 2002, vol. 4, p. 7.

Prabakaran U received his M.C.A degree in Computer Science in Shanmugha College of engineering, Thanjavur in 1999, Tamilnadu, India. Now he is doing his Master in Engineering in Vandayar College of engineering and technology, Thanjavur, Tamilnadu, India