2nd National Level Students' Research Conference on
"Innovative Ideas and Invention with Sustainability in Computer Science and IT-2021"
In association with International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

Tape Hardware Compression and Source Based Data Deduplication

Abhik Swarnakar[1], Rajesh Kumar[2], Anuradha Kanade*[2]
[1]Research Scholar, MCA(Science), SoCS , MIT-WPU, Pune, Maharashtra, India
[2]Assitant Professor, SoCS , MIT-WPU, Pune, Maharashtra, India

ABSTRACT

Cloud Computing has become an important advance for business actions in industry at the present time. These decades are experiencing the fast development of cloud computing that results in a vast data produced every moment. The data compression is becoming more important as it helps potentially in transportation over the network and efficient data storage to the great extent. This leads to the requirement of huge data processing and computation which is not easily accessible at the user's end. This has already led to the evolution of cloud platform for data storage. But solution to one problem may give birth to the other problem. Similarly, the speed of uploading and downloading the data from the cloud reduces the data processing time. Current paper focuses on providing solution to this problem by compressing the data using efficient tape hardware. It mainly considers the multimedia data for compression and also uses a hybrid technology for source-based cloud data deduplication for data stored on google drive and other networks.

Keywords – Data Compression, Data Deduplication, Hardware Compression, Tape Drive, Hybrid Algorithm, Virtual Tape Library (VTL)

I. INTRODUCTION

The cloud computing is a predictable field of data communication and resource sharing in the modern era. The real meaning of its functioning, its limits and the development of new applications, becoming increasingly agile and collaborative, inspiring subjects for research. The data decompression is nothing but to restore the compressed data back to its original form. It is also termed as expansion. The data compression is used for saving the resources like disk space, time needed to transmit or communicate the data over the internet. The data which is to be handled and the information to be communicated are growing. In this situation the data compression technology is considered as the important factor to handle the information. Hence it is essential to have an efficient algorithm to compress and decompress the humongous data present in the real world and to reduce the work load of the system. The primary objective of the proposed research is to find the Hybrid algorithm to optimize the resource consumption mentioned above.

In following section tape hardware compression technique is explained.

II. BACKGROUD WORK

Tape hardware compression is a technique in which data are been compressed and stored in the tape libraries. In tape hardware compression the data are been stored in a tape drive. Tape drives are an offline archival data storage which uses magnetic tape to store the data in a compressed manner. Tape drives are similar to a hard disk drive that provide direct access to the storage [2]. To read a particular piece of data, the Tape drives physically wind tape between reels. Tape drives have a very large access time. But the data can be streamed fast if a required position has been reached. Tape hardware compression if faster than the other compression technique because it doesn't slow things down. Choosing a tape hardware compression technique provide fast access of data compression. It has a better compression ratio of more than 2:1.[2] It uses Virtual Tape Libraries (VTLs) to compress the data and stores it on the Tape drive. It supports the transfer rate of up to 150MB/s. The IBM TS1160 has a capacity of 20TB. [2]

In the paper, "Hybrid Data Deduplication Technique in Cloud Computing for Cloud Storage", the authors have designed the hybrid data deduplication for cloud computing. It fulfills the demands of users as well as applications. They have used the file level and chunk level deduplication. Using hybrid design the researchers could achieve the effective data deduplication for various types of data. The FFCD design showed results closer to FVCD for chunk size of 500 bytes [8].

In the white paper by Oracle, the performance of the Tape Drive is evaluated for different size of the data. According to the paper, the speed of the storage applications or the throughput speed are the limiting factors in evaluating the performance of the tape drive. The existing storage application with the throughput of 50-60 MB/s is not sufficient to achieve the maximum speed of 4GB SCSI FCP interface. The current tape drive technologies are not capable to read/write the data for the 4GB SCSI FCP interface. Oracle mainly focused on doubling the throughput and considering the native drive performance on StorageTek T10000 tape drive. It proved very efficient in solving customer problems [9].

Tape drives have built in compression algorithms. Hardware compression might be less useful for the data which is secured by data protection operations. If sometime the network becomes jammed then the tape drives can become ravenous for data.

The hardware compression is reputable on the data path level. This kind of compression is only available for data paths which directly connects with the data to tape libraries. Though the compression structure the data sends uncompressed data from the client computer through the data path. The tape drive hardware compresses the data before writing it to the media.

Data compression automatically switches back on, when data becomes compressible again. Both the compressed and uncompressed data can be recorded on a tape which will be marked accordingly for proper treatment during playback. Such intellect prevents the expansion of early compressed or incompressible data.

Though the incompressible data is being used while the compression feature of a tape drive which does not have an intellect compression feature. This can cause a 5 - 10 percent lessening in capacity.

A mathematical algorithm is used that reduces terminated strings of data and assists in data compression. This in turn confirms the increased storage capacities of data. The compression algorithm is implemented using hardware for tape drives. It removes the terminated level from the data by encoding the pattern of input characters in efficient way. If the data patterns are repeated then the data compression will be more. The data deduplication can be achieved for such repeated patterns. However, the hardware compression technique is not much useful if the data secured by data protection

operations is competing with the other data used for the network bandwidth. In such case, the tape drive will compress the data but the data will not be abandoned quickly. The drives must start and stop the media as early as the data is available. Due to this the performance of the compression is affected.

The paper says that, the hardware compression is faster than the software compression as it is performed by dedicated electronic equipment. Mainly this compression is applicable for data paths. The data paths route the data to the tape libraries. Before data is written to the media, it is compressed by the tape drive. The hardware compression is used for direct-connect

configurations in which the sub-client and mediaagent are attached with the same physical computer. The data transfer to the media drives is smooth in these cases. Once the data is received from the sub-client it is quickly compressed by the drive. The tape can store the more data per unit time due to the high-speed operation of the tape.

The only drawback of the hardware compression is that it cannot be applied for the disk libraries. Therefore, the software compression is used for sub-client for the data paths associated with the libraries [11].

The compression feature of any software package must be turned off when the compression is available in the tape drive hardware or firmware. It helps to reduce the processing overhead on the computer system. The compression ratio 2:1 means that the compressed file is half of the size of the original file [10].

III. HARDWARE COMPRESSION

It accepts the uncompressed data (say image) from the client computer and is sent to the media through the data path. Before writing to the media, the tape drive hardware compresses the data (e. g. image).

Fig. 1: IBM TS1160 Tape Drive [3]

IV. SOURCE BASED PRIMARY DATA DEDUPLICATION

Primary data deduplication is a relatively recent trend in the field of file-based system storage solution as compared to backup-based data deduplication. There is only one copy named primary copy of the data for which no backup copy is available.

The main challenges for primary data deduplication are that the proposed system should be able to balance the system resource consumption (CPU/memory/disk I/O) with the deduplication space savings and the deduplication throughput i.e., the speed with which the data is being send to the cloud server after compression [2].

Following are the main requirements for the system that are identified.

A) Optimize for the Uniqueness of the Data

The optimization for the uniqueness of the data can be achieved using hardware compression. The bulk of the data can be unique on the contrary to the backup-based data deduplication where 90% of the data is duplicated.

B) Broadly Used Platform

Specifically, it must run hassle free and efficiently on a broadly used platform like the windows server 2012.

C) Minimum Requirements

Particularly, it must run-on low-end servers and should be able to accommodate huge variations in workloads and different hardware platforms.

D) Friendly Primary Workload

The proposed system cannot assume that it will be having access to dedicated resources and hence must yield to primary workload.

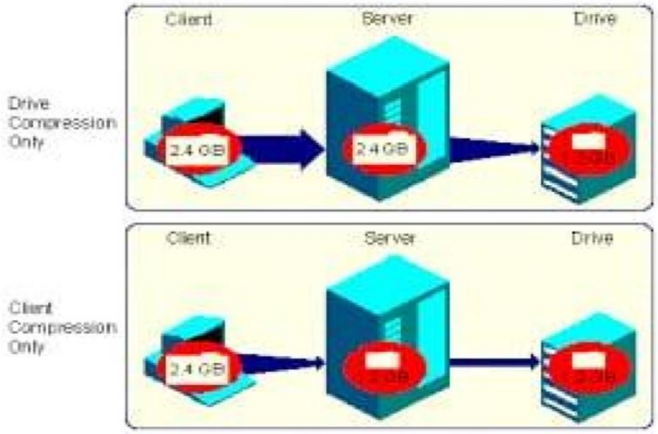

Fig. 2: Tape volume capacity and data compression.

V. MATHEMATICAL EQUATIONS

Usually, the hardware compression is much faster than a software compression because as disparate to the software compression it does not use a computer processor which draws from resources. The following is the illustration for the same.

If we have 200 MB of data under existing system.

(i) Using Tape Hardware Compression Alone:

If the compression ratio is 2:1 then 200 MB data is compressed to 100 MB.

(ii) Using Source Based Data Deduplication Alone:

If the compression ratio is 3:2 then 200 MB data is compressed to 133.3 MB.

Under proposed system first we use DATA DEDUPLICATION at software level so 200 MB gets converted to say 134 MB. Then, we use TAPE HARDWARE COMPRESSION at the hardware level so 134 MB gets converted to 67 MB.

Therefore, Compression Ratio Achieved using the proposed Hybrid Algorithm is, available in the tape drive hardware. The compression features the software package which should be turned off. This reduces the processing load on your computer and also the hardware-based compression which is naturally much more effective.

VI. VIRTUAL TAPE LIBRARY (VTL)

A **virtual tape library** (**VTL**) is a data storage technology used in general for backup as well as recovery purposes. It presents a logical view of the physical storage resources to the host computer. [1]

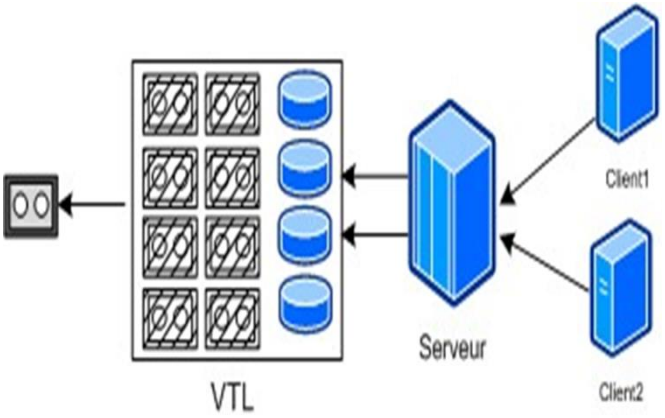

Fig.3: Virtual Tape Libraries (VTLs) [4]

VII. CONCLUSION

In this proposed work, we are making a Hybrid technology using tape hardware compression and source-based data deduplication. Source based data deduplication compresses the data (i.e., 2:1 ratio) and then it interacts with the tape hardware system to store that compressed data into more compressible manner (i.e., 4:1 ratio). We are combining the hardware compression and the software compression to obtain a lossless form of data upload and unload. We are using tape hardware so that the amount of time taken in uploading and retrieving the data becomes faster. We are working on the algorithm how to connect the two data compression technique and form a hybrid of it. So that this technology may be useful in future.

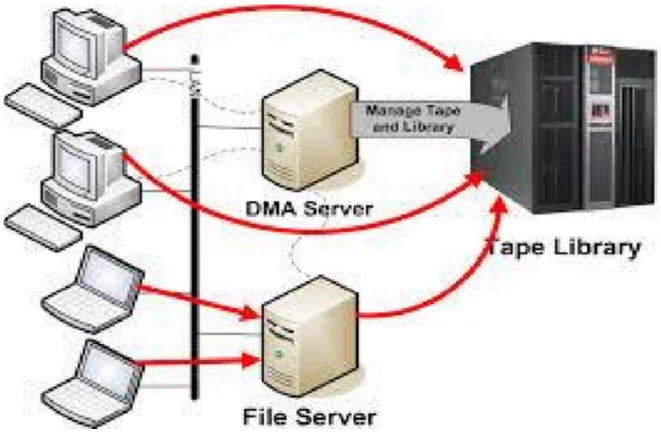

Fig. 5: Checking Tape Driver Performance

VIII. REFERENCES

[1]. "Information Storage and Management", EMC Education Services, (2010), John Wiley & Sons.p.210, ISBN 978-0-470-29421-5, Retrieved February 16, 2021

[2]. https://en.wikipedia.org/wiki/Tape_drive, retrieved on February 19, 2021

[3]. https://www.ibm.com/products/ts1160, accessed on February 15, 2021

[4]. https://commons.wikimedia.org/wiki/File:Virtual_Tape_Library.png, accessed on February 15, 2021

[5]. D. T. Meyer and W. J. Bolosky, "A Study of Practical Deduplication", 2012, [online] Available: http://static.usenix.Org/.

[6]. K. Jin and E. L. Miller, "Deduplication on Virtual Machine Disk Images", 2010.

[7]. Daehee Kim, Sejun Song and Baek Young Choi, "SAFE: Structure-Aware File and Email Deduplication for Cloud-based Storage Systems", IEEE, pp. 130-137, 2013.

[8]. "Hybrid Data Deduplication Technique in Cloud Computing for Cloud Storage", 2017 Journal of Theoretical and Applied Information Technology, Vol. 95 No. 24, pp. 7069-7081

[9]. Dwayne Edling, "Evaluating Tape Drive Performance", 2011 Oracle, Sun Storagetek, pp. 1 -7

[10]. "Flash Technical Support", online: https://asset.fujifilm.com/www/us/files/2020-03/417ca5e652a2068f3918a9781bb75915/Data_Compression_FAQ_01-11.pdf, Accessed on February 18, 2021. [11]. "Use Tape Drive Compression to Save Storage Space", 2019, online: https://documentation.commvault.com/commvault/v11/article?p=12329.htm, accessed on February 12, 2021

[11]. https://www.ibm.com/support/pages/sites/default/files/inline-files/IBM%20Data%20Deduplication.pdf , accessed on January 3, 2021.