International Journal of Recent Trends in Engineering & Research
ISSN (ONLINE) : 2455 – 1457
IMPACT FACTOR : 4.101
ICNSCET19- International Conference on New Scientific Creations in Engineering and Technology

A SURVEY ON DEDUPLICATION TECHNIQUES FOR STORAGE OPTIMIZATION

Pavithra M[1], Jane Rubel Angelina J[2]
[1] *Department of Computer Science and Engineering, Thiagarajar College of Engineering,*
[2] *Department of Computer Science and Engineering, Thiagarajar College of Engineering,*

Abstract— In the Technical world, there is a lot of data increased day by day. These data like text, image, audio, video, etc. These are creating many issues in both the storage and retrieval process. The organization spent the most amount of to cost to retrieve a backup copy. Hence a capable method is to be introduced to handling a large amount of data. Some accessible methods available for removing the repeated data in a backup system are Deduplication and Data Reduction. Data deduplication is the specialized form of data compression technique which eliminates the repeated data, minimize the amount of bandwidth utilization and also maximize the storage space and cost. A mixture of research papers has been collected from the literature study, as the outcome of this paper endeavors to review the techniques of backup storage optimization ideas and methods using deduplication and also the chunking process should be reviewed by this Deduplication Process.

Keywords— Deduplication, chunking, data compression, Disk space, Backup speed

I. INTRODUCTION

In [1] Recent years the availability of data should be increased so all over world approximately 1.8 zettabytes data were produced and pretend that it will be exciting to 7.9 Zettabytes by the end of 2015 this should be estimated by International Data Corporation(IDC). In global backup servers will be escalated in more than 10 times over the next era. In 2020 there are lots of data from social media, organization, digital information, sensors gadgets, construction design like interior and exterior structures of buildings will full-fledged by more than 50 times. Most industries are a tussle with the decay of data and also a concern with the protection process. Thus most of the industries want the solution to overcome this growth of data. The lots of information should be stored in terabytes, terabytes with the help of Data warehouse. This stored information has lots of repeated data so there is a coincidental to store the information within the backup system. Even though keeping these enormous amount of data it may impinge on the action, bandwidth, storage space consistencies and so on., for example we put the common byte pattern possible arise many times, assume that the number of time that make some small changes in word file or other assets files the volume of duplicate can be considered at the time. In some organization, 80% of corporate data information is

duplicate across the organization. To indicate that issue that has many techniques like data compression, data deduplication can be used to improve storage space effectively thereby removing repeated data. Replacing the volume of information forward through the network can save more money, storage space, the speed of recovery in backup-in various cases reserves up to 90 %. Nowadays many cloud service providers provide a free storage space of private data. Example Dropbox provides 2GB storage space for private data and Recently Google search engine enterprise introduce Google Drive which also affords free online storage space for 5GB. These organizations have an efficient deduplication file system to effectively managing redundant data. There are some storage solution providers such as EMC and Netsapp introduce data deduplication software for backup storage. This software is used to minimize the amount of data everyday transfer from user to backup server through the internet at the time minimizes the cost and bandwidth. These papers have the following contribution: Section 2 summarizes various optimization techniques in storage, section 3 explains introduction about deduplication, basic for deduplication, various Techniques used to implement the deduplication and Section 4 Conclusion of this paper.

II. STORAGE OPTIMIZATION TECHNIQUES

To store data, we use several kinds of storage optimization methods. They are Thin provisioning, Snapshots, Clones, Deduplication and Compression. Data deduplication and data compression are more commonly used methods. These techniques are explained in sections 3 and 4.

2.1. Thin Provisioning

In recent times, a technology called thin provisioning is used in storage optimization. When applications run in out of storage, this shared-storage environment relies on on-demand allocation of blocks of data to reduce the risk of application failures. Almost, all the white spaces are eliminated by this methodology. This helps to avoid poor utilization rates and to acquire higher storage utilization [2]. Thin provisioning system eliminates the unused capacity of a physical disk to achieve higher storage utilization. Hence, it is recognized as one of the best technique. The unused allocated storage cannot be used by any other applications while using traditional allocation and thin provisioning. Moreover, the full volume of storage is never used but it is essentially wasted. This condition is sometimes called stranded storage. Usually, to store an exact amount of data as per the need and to remove the wasted storage capacity on paying, thin provisioning methodology is preferred [3]. In addition to this, an additional volume can be added to the existing combined storage system when we need more storage. Thin provisioning storage space on an on-demand basis. For example, in Gmail, every Gmail account contains a large amount of allocated capacity but most of the Gmail users use less amount of the allocated storage space. In this case, the thin provisioning method will be highly useful for eliminating the unused space.

2.2. Snapshot Technology

In Snapshot technology, the user is allowed to store only the changes between each dataset that are frequently accessed multiple times for various reasons. The snapshot technologies are used at the operating system level by some storage vendors to enable and access data in application level layers. The term "clones and snapshots" looks confusing at present because the care should be taken when evaluating the vendor claims. Full point-in-time copies of "snapshots" or "clones" are used by some vendors [3], and some use the same term to refer shared-block "delta" snapshots or clones. For an implementation of reading only snapshots, some vendors use this technology while others use this to provide writable ones. This technique is formally known as "delta snapshot" technology. A sheet of paper containing complete contact information is sent to all the authors. This includes full mailing addresses, telephone numbers, fax numbers, and email addresses. Using this information, each author is sent a complimentary copy of the journal in which the paper appears. In addition, designate one

author as the 162 "corresponding author". The proofs of the paper will be sent to this author. Proofs are sent to the corresponding author only.

2.3. Clones

An advanced form of writable snapshots is generally known as clones. They are essentially a snapshot volume presented as a 'real' volume can be modified or changed. Initially, clones had Clones are primarily used for test and development of applications. Because of this reason, they have limited value. In reducing the storage footprint, clones have immense values in various environments, as the virtualization field, especially the desktop virtualization, rises every day [3]. Since hundreds of virtual machine-based storage images can now be loaded into a cache, the performance can be improved.

2.4. Data Deduplication

Data deduplication is a technique which is used to track and eliminate the duplicate chunks (piece of data) in a storage unit. This technology is used by many vendors to implement efficient data storage. There are separate merits and demerits in this technique. Deduplication is more important at the shared storage level [4]. It is important implementations in software, as well as the database. Platform virtualization and backup server are the most suitable candidates for deduplication because both applications will use and produce a lot of identical/duplicate copies. In-place deduplication deduplicates primary storage and is offered by a few vendors in recent times. Deduplication takes place on both the file level and block level storage system. The duplicate or redundant copies in the same file are eliminated while using file-level deduplication method. This type of deduplication is known as single instance storage (SIS). Similarly, the redundant or duplicated blocks of data which is present in unique files are eliminated, while using block level deduplication [5]. Block-level deduplication reduces more space than SIS [6]. This type of deduplication is known as a variable block or variable length deduplication. The word data deduplication is used as a synonym for block-level deduplication often. Sometimes, it is mentioned as variable length deduplication. This is explained in detail in section 4.

2.5. Data compression

Compression is a mechanism which is used to save the storage space. This is done by removing the binary level redundant data within a data block [7]. Unlike deduplication, in compression technique, the data are always stored only in the most efficient block. In this storage method, it is not concerned with the choice of the second copy of the same block exists or not. The requirement of memory source is relatively small because the compression technique works within a data block and also only one file at a time can be looked at. [3]. some of the day-to-day live examples for file level compression technique are JPG image, audio and video files.

III. DATA DEDUPLICATION

Data deduplication is a special form of well grown optimized method to store and minimize the amount of investment cost of the organization by reducing the storage space and bandwidth cost. It is very helpful for cloud service providers because this method wants minimum hardware requirements to save the data. The benefits of data deduplication are listed below [8]

1) Needs low hardware cost for implementation
2) Low backup cost
3) Maximize the storage space
4) Increase the network speed efficiently

5) Reduce the bandwidth while transferring data

Data deduplication is the method of removing repeated data [9].In this technique, the exact copy of data is removed and the common chunks(sequence of bytes) of data, or patterns which are detected and stored during this process and minimize the disk space and also reduce the bandwidth while transferring the data through the network. In this experiment, the chinking blocks are compared to every block in the disk storage, at the time an exact copy of duplicates are found that block is replaced with reference pointer at that pointer place used to refer the original block of data in the disk. For example, the same block of data arise more times during the process but the match size depends on chunk size, by using this frequency match the data must be sent and stored with less space for storing that matched data.

3.1. Working Principles of Deduplication

Basic terms of deduplication are the process of removing repeated data files or block and compare each and every block in the existing Storage [10]. Deduplication removes an exact copy of a given block in the storage space. The deduplication process consists of the following steps:

1)Divide the input data into blocks or chunks.

2) Generate the fingerprint (hash values) using any hash function for every chunk of data.

3) Using fingerprint value check whether the fingerprint of data is present in any of the existing block of data.

4) If the exact copy of data is found the reference pointer should be given to the newly created fingerprints in the database.

5) Finally, the results show the unique chunks are stored in storage otherwise duplicates are eliminated.

IV. IMPLEMENTATION METHOD OF DEDUPLICATION

There are several methods for eliminating redundant data in the storage space. Still, most of the industries or organization use the data deduplication techniques to remove the duplicate data and minimize the redundant problem [4].the following methods done by deduplication there are

A) Location-based deduplication

B) Time-based deduplication

C) Chunk-based deduplication

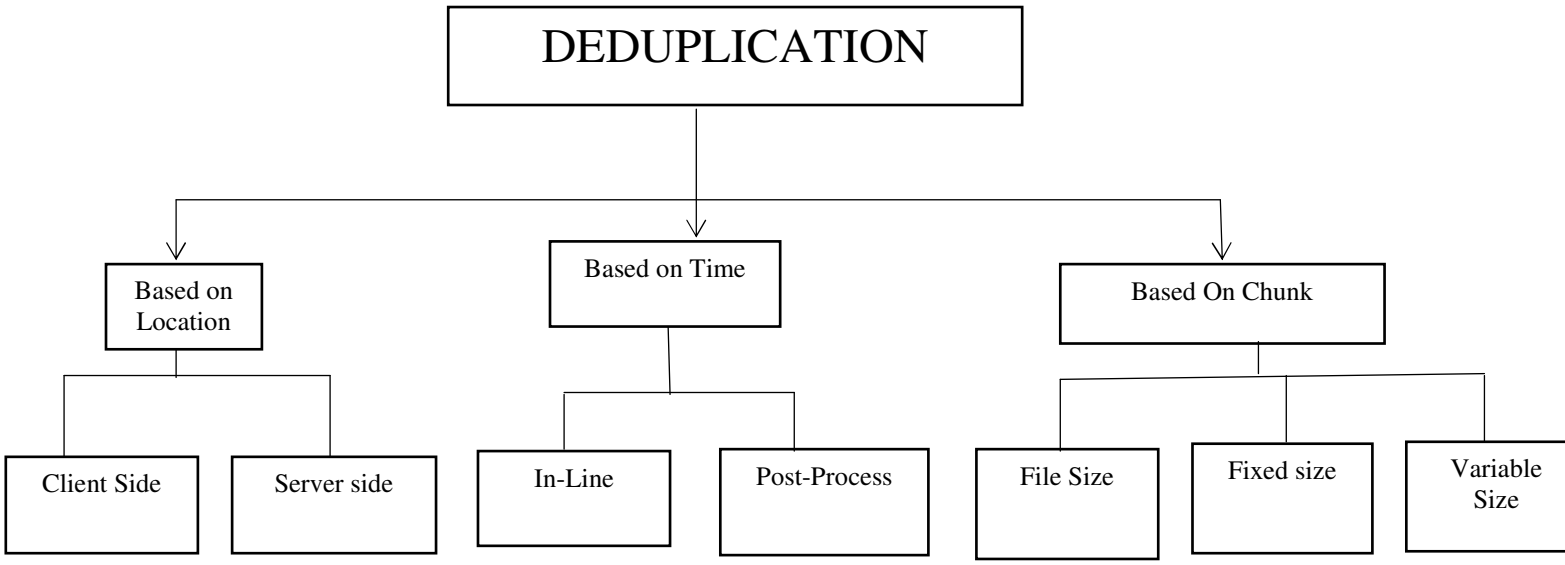

Fig 1.Types of Deduplication techniques

4.1. Location based Deduplication

In this process, deduplication is performed at various locations. Based on this location the entire process of deduplication has performed that place
Deduplication process can be performed at different locations. Based on which location, the deduplication process is happening; the entire deduplication is carried out either at the source side (Client) or target side (Server) [8].

4.1.1. Client-side Deduplication

The client-side deduplication is done in the user side or source side. Fig. 3 explains that the client side deduplication process. The based client deduplication process is done by place a dedupe engine at the physical or virtual server. At that time dedupe engine will detect all the duplicates data over the backup server and also transmit the unique data blocks to the disk. All this process is done before transmitting the data over the network. The benefit of client-side deduplication is it consumes minimum bandwidth and also modified data gets backed up.

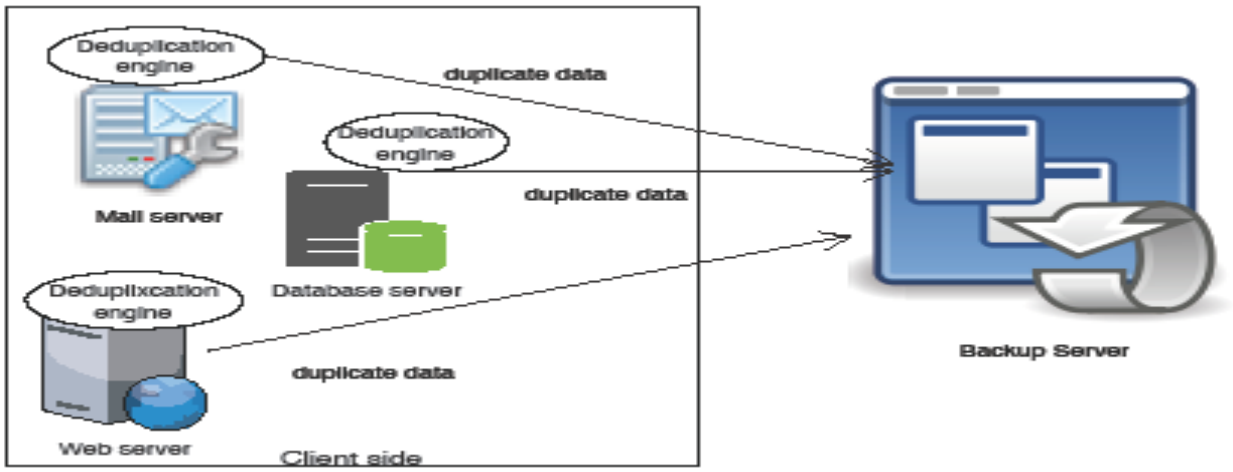

Fig 2.Client side deduplication

4.1.2. Server-side Deduplication

In server-side deduplication [8], is done in a target server or destination server. Fig 3 explains the deduplication perform server-side appliance. At that time dedupe engine check the data after receiving the backup data from the client. The benefit of this method is to reduce the overhead of the client elimination process.

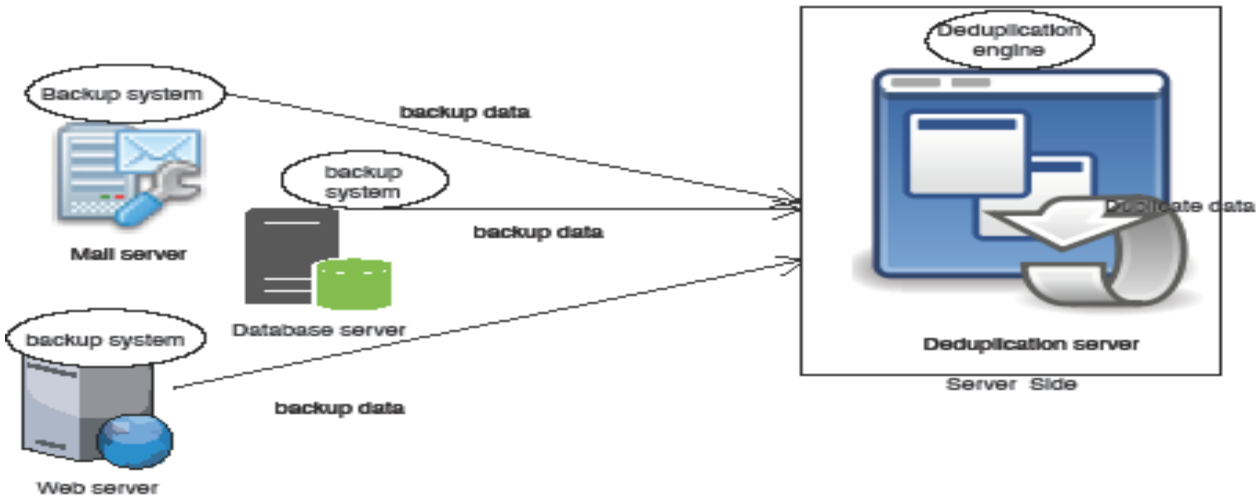

Fig 3.Server side deduplication

4.2. Time-Based Deduplication

In this method deduplication process based on "which time to duplicate the data". Data can be processed in three ways the before being written into the disk (Inline) and after writing to the disk (Post), or both before and after writing to the disk (Hybrid) [4].

4.2.1. In line deduplication

The data is transfer from client or source to the Server or Target. It can only be done on the client side. An Inline deduplication process data is deduplicated before it is written to disk. If a block of data arrives into the process it checks whether the block of data has been processed already or not. Before processing this data it was dragged from redundant block then put a reference point to the block. If it found that block was unique to existing block the process write the block into the Storage. While investigating data block is initiated before it is written. In this technique used in RAM to minimize the overhead in 110 and also saves storage space. Even though it requires large resources and this becomes a network's bottleneck [9].The benefits of Inline deduplication is it doesn't require additional disk space. Few companies like Hewlett packed, Diligent Technologies offers Inline and Chunk-based deduplication Products [10], [11] [12].

4.2.2. Post process Deduplication

In this post, process deduplication is performed in the target side or server side. The client data is written to the backup storage and the duplicate data are cleared after the backup process. In this process started once data is written into storage space. At that time the process was started and retrieves the data space [13]. The Benefit of Post-process deduplication is that the performance is maximum than In-line process deduplication and also this method is used to share index and metadata. In the post-process deduplication high availability clustering is easier, and replicating data can be more efficient. The drawbacks of this process need fast disk cache, which usually makes more cost to initialization compare than Inline process deduplication. However since non-deduplicate data does not need to be stored, in this process takes a long time to process this duplicates may be more costly in long run.falcon Stor FDS and SeptonDeltaStor, Exagrid this company offers Post process Deduplication[10][11][12].

4.3. Chunk based deduplication

In this Chunk-based deduplication the data are spiltted into a sequence of bytes called Chunks. These chunks are used to look at the redundant bytes. This Chunking process split the data into the number of pieces called chunks or blocks. Only the unique chunks are stored in the disk. The various chunk-based deduplication are available in the survey to remove the redundant data which is present in the disk space or backup system

1) Single Instance Storage or whole file chunking
2) Fixed Size Chunking
3) Variable Size Chunking

4.3.1 Single Instance Storage (SIS) or whole file Chunking

In this SIS Chunking, the files cannot be split into chunks rather than it collected small files and consider these files as a whole file as a chunk. It generates a hash value for the entire chunk is called file index, then it is checked whether this index to existing file index in the disk. If a new incoming file matches with file index, then it is assumed as duplicate and put the reference pointer to duplicate file. This reference is used to find the existing index in the system.

4.3.2. Fixed Size Chunking

This fixed size chunking the files are split into fixed or equal sized chunks in which the chunk size is 4KB,8KB,16Kb, etc., for example, we fix that the chunk size is 4KB a file is chunked at 8KB,10KB,20KB continuously. As a result the content based checksum identified the chunks and store only the index which does not exist [13]. This fixed size chunking overcomes this issue in SIS. Assume that large file which is modified only a few bytes at that time the chunks are reindexed and stored in the backup space. For example, we take 5GB document which is changed by the user in 100kb. The existing document and new document have different checksums but SIS stores the full document and results 10GB. Otherwise using Fixed size chunking the calculate the document size like this way (changed file size/Existing file size)*Existing file size. For this example, we get only 104KB of data for a given document is stored.

4.3.3. Variable size chunking

In this process vary from fixed size chunking. Here, chunking boundaries are determined based on the contents of the file. Thus it is tougher to insertion and deletion. Nowadays this has been thought of as the finest algorithm for the backup system [14]. Like Fixed size chunking method a variable size chunking has the following steps,

i) Split the Files into variable blocks based on the chunk size.
ii) Generate the hash values forever block in the file.
iii) Check whether any redundant data from generated hash values.
iv) This algorithm which is used for finding chunking boundaries is the Rabin fingerprinting algorithm.
v) Every chunk converted into hash values using common techniques like MD% or SHA-1

V. COMPARISION METRICS OF DIFFERENT DEDUPLICATION METHODS

Table 1, shows the various metrics which has been measured across the deduplication tactics such as file level, fixed size, variable size deduplication ratio, processing time, Index overhead. The deduplication ration is nothing but the value of a given amount of space saved

during the deduplication process. It is the comparison of bytes before and after deduplication [17]. Fig 4 shows that the comparison results of fixed and variable sized chunking methods. T hose are taken based on deduplication ratio. For the evaluation of 10kb,100kb, 1Mb capacity file has been taken has to be taken and taken the results. Deduplication would work better for large datasets.

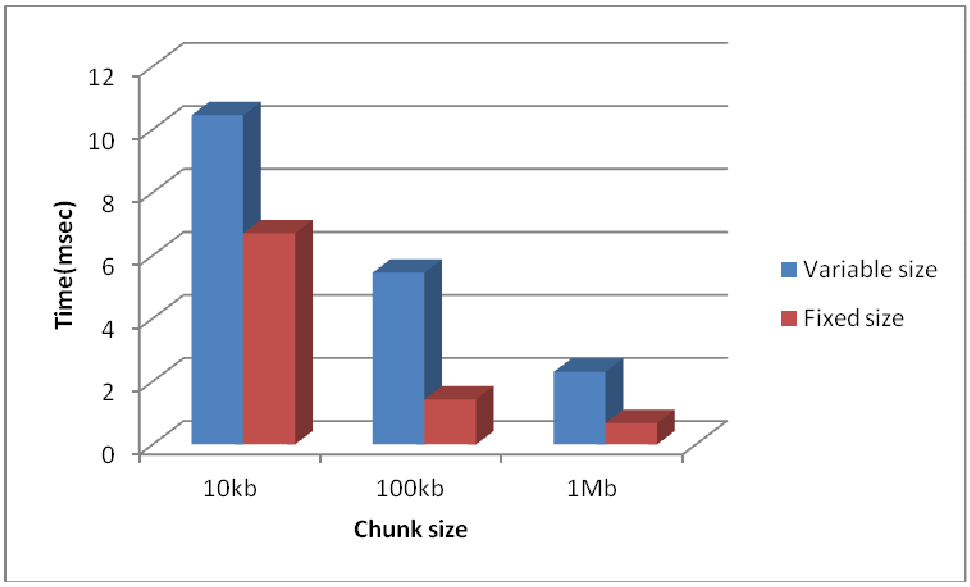

Figure 4: Comparison of Fixed size versus Variable Size

Metrics	File Size	Fixed Size	Variable Size
Deduplication ratio	Good	Better	Best
Processing Speed	Medium	low	High

Table 1: Comparison of performance matrices across file size, fixed size, Variable size Deduplication Methods

It has been improved that file level works better in client side and variable size works better in server side [24].

VI. CONCLUSION

In this paper, we have to survey the different type of deduplication techniques. In Between, we determine that Variable size data deduplication is well-performed compare than the other types by comparing the hash of each chunk in the given document. In the future, this method increases the storage space and so increases the performance by supporting storage resources to transfer and took more data. In the future, more research works under variable size chunking to increase the performance, reduce the processing time and improve the large scale data storage. Then to develop a proficient method to minimize the fragmentation and get high read and write throughput.

REFERENCES

[1] Walid Mohamed Aly, Hany AtefKelleny, "Adaptation of Cuckoo Search for Documents Clustering," International Journal of Computer Applications (0975 - 8887), Volume 86 – No-1,2014.

[2] Xian Chen, Wenzhi Chen, Zhongyong Lu, Peng Long, Shuiqiao Yang, Zonghui Wang, A Duplication-Aware SSD-Based Cache Architecture for Primary Storage in Virtualization Environment, IEEE Systems Journal, Volume: 11, issue:4, Dec. 2017

[3] Jianwei Yin, Yan Tang, Shuiguang Deng, Ying Li, and Albert Y. Zomaya, Fellow, D3: A Dynamic Dual-Phase Deduplication Framework for Distributed Primary Storage, IEEE Transactions On Computers, VOL. XX, NO. XX, JULY 2017

[4] Philipp C. Heckel (2013, May 20). "Minimizing remote storage usage and synchronization time using deduplication and multi chunking," [Online]. Available: http://blog.philippheckel.com!

[5] Xuecheng Zhang, Mingzhu Deng, An Overview on Data Deduplication Techniques, Information technology, and intelligent systems pp -359-369 2016

[6] Naresh Kumar,Preeti Malik, Sonam Bhardwaj, Sushil Chandra Jain, "Comparative analysis of deduplication techniques for enhancing storage space",2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)

[7] Sandip Agarwal a, Divyesh Jadav, Luis A Bathen, "iCostale: AdaptiveCost Optimization for Storage Clouds," IEEE 4th International Conference on Cloud Computing, IEEE, 2011

[8] Chris Poelker (Aug 20, 2013). Intelligent Storage Networking [Online].Available: http://www.computerworld.com/

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

[10] Benjamin Zhu, Kai Li, and Hugo Patterson, "Avoiding the Disk Bottleneck in the Data Domain Deduplication File System," Proc. of the USENIX File And Storage Technologies, 2008.

[11] Data Domain LLC. Data Domain Boost Software. [Online]. Available:http://www.datadomain.com/

[12] Symantec Corporation. Symantec NetBackup PureDisk. [Online]. Available:http://www.symantec.com/

[13] ExaGrid Systems. ExaGrid EX Series Product Line.[Online]. Available: http://www.exagrid.com/

[14] Wen Xia , Hong Jiang, Dan Feng, Fred Douglis, Philip Shilane, ,Yu Hua, Min Fu, Yucheng Zhang, Yukun Zhou," A Comprehensive Study of the Past, Present, and Future of Data Deduplication", Proceeding of the IEEE Volume:104 ,issue:9 , Sept. 2016

[15] Deepavali Bhagwat, Kave Eshghi, Darrell D. E. Long, and Mark Lillibridge, "Ex-treme binning: Scalable, parallel deduplication for chunk-based File backup," In MASCOTS, pp 1-9, IEEE, 2009.

[16] Jin-Yong Ha, Young-Sik Lee, Jin-Soo Kim, "Deduplication with Block LevelContent-Aware Chunking for Solid State Drives (SSDs), "HighPerformance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), pp 1982 - 1989,2013.

[17] M. Dutch, "Understanding data deduplication ratios," In SNIA Data Management Forum, 2008.

[18] George Crump (2011, September 30). Which Primary StorageOptimization is Best? [Online]. Available: http://www.storageswitzerland.

[19] D. T. Meyer, W. 1. Bolosky (2012), "A Study of PracticalDeduplication,"[Online]. Available:http://static.usenix.Orgf

[20] Min Li, Shravan Gaonkar, Ali R. Butt, Deepak Kenchammana, and Kaladhar Voruganti, "Cooperative Storage-Level Deduplication for Reduction in Virtualized Data Centers," IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems, pp.209-218, 2012.

